DOCUMENT RESUME

ED 417 451 CS 509 793

AUTHOR	Pisoni, David B.
TITLE	Research on Spoken Language Processing. Progress Report No. 21 (1996-1997).
INSTITUTION	Indiana Univ., Bloomington. Dept. of Psychology.
SPONS AGENCY	National Institutes of Health (DHHS), Bethesda, MD.
PUB DATE	1997-00-00
NOTE	637p.
CONTRACT	DC-00111; DC-00012
PUB TYPE	Collected Works - General (020) -- Reports - Research (143)
EDRS PRICE	MF03/PC26 Plus Postage.
DESCRIPTORS	*Deafness; Error Analysis (Language); Higher Education; *Language Processing; Memory; *Speech Communication; Word Recognition
IDENTIFIERS	Cochlear Implants; *Indiana University Bloomington; *Speech Perception; Speech Research

ABSTRACT

This 21st annual progress report summarizes research activities on speech perception and spoken language processing carried out in the Speech Research Laboratory, Department of Psychology, Indiana University in Bloomington. As with previous reports, the goal is to summarize accomplishments during 1996 and 1997 and make them readily available. Some papers in the report are extended manuscripts that have been prepared for formal publication, others are short reports of research. The extended manuscripts are: "Speech Perception" (R. Wright and others); "Looking at the 'Stars': A First Report on the Intercorrelations among Measures of Speech Perception, Intelligibility and Language Development in Pediatric Cochlear Implant Users" (D.B. Pisoni and others); "Measures of Phonological Memory Span for Sounds Differing in Discriminability: Some Preliminary Findings" (M. Cleary); "Static vs. Dynamic Faces as Retrieval Cues in Recognition of Spoken Words" (L. Lachs); "Some Observations on Working Memory Tasks and Issues in Cognitive Psychology" (W.D. Goh); "Familiarity, Similarity and Memory for Speech Events" (S.M. Sheffert and R.M. Shiffrin); "Improvements in Speech Perception in Prelingually-Deafened Children: Effects of Device, Communication Mode, and Chronological Age" (T.A. Meyer and others); "Predicting Open-Set Spoken Word Recognition Performance from Feature Identification Scores in Pediatric Cochlear Implant Users: A Preliminary Analysis" (S. Frisch and D.B. Pisoni); "Lexical Competition in Spoken English Words" (S. Amano); and "Some Computational Analyses of the PBK Test: Effects of Frequency and Lexical Density on Spoken Word Recognition" (T.A. Meyer and D.B. Pisoni). Among the short reports are "Cognitive Factors and Cochlear Implants: An Overview of the Role of Perception Attention, Learning and Memory in Speech Perception" (D.B. Pisoni); "Performance of Normal-Hearing Children on Open-Set Speech Perception Tests" (M. Kluck and others); "Effects of Talker, Rate and Amplitude Variation on Recognition Memory for Spoken Words" (A.R. Bradlow and others); "Acoustic, Psychometric and Lexical Neighborhood Properties of the Spondaic Words: A Computational Analysis of Speech Discrimination Scores" (T.A. Meyer and others); "Effects of Alcohol on the Production of Words in Context: A First Report" (S.B. Chin and others); "Intelligibility of Normal Speech II: Analysis of Transcription Errors" (A.T. Neel and others); "Some Observations on Neighborhood Statistics of Spoken English Words" (S. Amano); "Sensory Aid and Word Position Effects on

+++++ ED417451 Has Multi-page SFR---Level=1 +++++

Consonant Feature Production by Children with Profound Hearing Impairment" (S.B. Chin and others); "Lexical Competition and Reduction in Speech: A Preliminary Report" (R. Wright); "A Preliminary Acoustic Study of Errors in Speech Production" (S. Frisch and R. Wright); "Experience with Sinewave Speech and the Recognition of Sinewave Voices" (S.M. Sheffert and others); "Some Factors Affecting Recognition of Spoken Words by Normal Hearing Adults" (A.R. Bradlow and others); and "The Hoosier Audiovisual Multi-Talker Database" (S. Sheffert and others). Contains references, data tables, and a publications list. (RS)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 417 451

RESEARCH ON SPOKEN LANGUAGE PROCESSING

Progress Report No. 21
(1996-1997)

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana
47405

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

- Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Supported by:

Department of Health and Human Services
U.S. Public Health Service

National Institutes of Health
Research Grant No. DC-00111
and
National Institutes of Health
Training Grant No. DC-00012

CS 509 793

BEST COPY AVAILABLE

RESEARCH ON SPOKEN LANGUAGE PROCESSING

Progress Report No. 21
(1996-1997)

David B. Pisoni, Ph.D.
Principal Investigator

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405-1301

Research Supported by:

Department of Health and Human Services
U.S. Public Health Service

National Institutes of Health
Research Grant No. DC-00111

and

National Institutes of Health
Training Grant No. DC-00012

1997
Indiana University

RESEARCH ON SPOKEN LANGUAGE PROCESSING

Progress Report No. 21 (1996-1997)

Table of Contents

Introduction vii

Speech Research Laboratory Faculty, Staff, and Technical Personnel viii

I. Extended Manuscripts xi

- Speech Perception
  Richard Wright, Stefan Frisch, and David B. Pisoni 1

- Looking at the "Stars": A First Report on the Intercorrelations Among Measures of Speech Perception, Intelligibility and Language Development in Pediatric Cochlear Implant Users
  David B. Pisoni, Mario Svirsky, Karen I. Kirk, and Richard T. Miyamoto 51

- Measures of Phonological Memory Span for Sounds Differing in Discriminability: Some Preliminary Findings
  Miranda Cleary 93

- Static vs. Dynamic Faces as Retrieval Cues in Recognition of Spoken Words
  Lorin Lachs 141

- Some Observations on Working Memory Tasks and Issues in Cognitive Psychology
  Winston D. Goh 179

- Familiarity, Similarity and Memory for Speech Events
  Sonya M. Sheffert and Richard M. Shiffrin 201

- Improvements in Speech Perception in Prelingually-Deafened Children: Effects of Device, Communication Mode, and Chronological Age
  Ted A. Meyer, Mario A. Svirsky, Karen I. Kirk, and Richard T. Miyamoto 235

- Predicting Open-Set Spoken Word Recognition Performance from Feature Identification Scores in Pediatric Cochlear Implant Users: A Preliminary Analysis
  Stefan Frisch and David B. Pisoni 261

- Lexical Competition in Spoken English Words
  Shigeaki Amano 289

- Some Computational Analyses of the PBK Test: Effects of Frequency and Lexical Density on Spoken Word Recognition
  Ted A. Meyer and David B. Pisoni 315

II. Short Reports and Work-in Progress 333

- Cognitive Factors and Cochlear Implants: An Overview of the Role of Perception Attention, Learning and Memory in Speech Perception
  David B. Pisoni 335

- Performance of Normal-Hearing Children on Open-Set Speech Perception Tests
  Melissa Kluck, David B. Pisoni, and Karen Iler Kirk 349

- Effects of Talker, Rate and Amplitude Variation on Recognition Memory for Spoken Words
  Ann R. Bradlow, Lynne C. Nygaard, and David B. Pisoni 367

- Acoustic, Psychometric and Lexical Neighborhood Properties of the Spondaic Words: A Computational Analysis of Speech Discrimination Scores
  Ted A. Meyer, David B. Pisoni, Paul A. Luce, and Robert C. Bilger 385

- Effects of Alcohol on the Production of Words in Context: A First Report
  Steven B. Chin, Nathan R. Large, and David B. Pisoni 403

- Intelligibility of Normal Speech II: Analysis of Transcription Errors
  Amy T. Neel, Ann R. Bradlow, and David B. Pisoni 421

- Some Observations on Neighborhood Statistics of Spoken English Words
  Shigeaki Amano 439

- Sensory Aid and Word Position Effects on Consonant Feature Production by Children with Profound Hearing Impairment
  Steven B. Chin, Karen Iler Kirk, and Mario A. Svirsky 455

- Lexical Competition and Reduction in Speech: A Preliminary Report
  Richard Wright 471

- Training Japanese Listeners to Identify English /r/ and /l/: Long-Term Retention of Learning in Perception and Production
  Ann R. Bradlow, Reiko Akahane-Yamada, David B. Pisoni, and Yoh'ichi Tohkura 487

- A Preliminary Acoustic Study of Errors in Speech Production
  Stefan Frisch and Richard Wright 503

- Experimental Evidence for Abstract Phonotactic Constraints
  Stefan Frisch and Bushra Zawaydeh 517

- Experience with Sinewave Speech and the Recognition of Sinewave Voices
  Sonya M. Sheffert, David B. Pisoni, Nathan R. Large, Jennifer M. Fellowes, and Robert E. Remez 531

- Tongue Twisters Reveal Neighborhood Density Effects in Speech Production
  Michael S. Vitevitch 545

- Some Factors Affecting Recognition of Spoken Words by Normal Hearing Adults
  Ann R. Bradlow, Gina M. Torretta, and David B. Pisoni 557

- Audio-Visual Speech Perception Without Traditional Speech Cues: A Second Report
  Robert E. Remez, Jennifer M. Fellowes, David B. Pisoni,
  Winston D. Goh, and Philip E. Rubin 567

III. Instrumentation and Software 575

- The Hoosier Audiovisual Multi-Talker Database
  Sonya Sheffert, Lorin Lachs, and Luis Hernández 577

IV. Publications: 1996-1997 585

INTRODUCTION

This is the twenty-first annual progress report summarizing research activities on speech perception and spoken language processing carried out in the Speech Research Laboratory, Department of Psychology, Indiana University in Bloomington. As with previous reports, our main goal has been to summarize our accomplishments over the past two years and make them readily available to granting agencies, sponsors and interested colleagues in the field. Some of the papers contained in this report are extended manuscripts that have been prepared for formal publication as journal articles or book chapters. Other papers are simply short reports of research presented at professional meetings during the past two years or brief summaries of "on-going" research projects in the laboratory. From time to time, we also have included new information on instrumentation and software developments when we think this information would be of interest or help to others. We have found the sharing of this information to be very useful in facilitating research.

We are distributing progress reports of our research activities because of the ever increasing lag in journal publications and the resulting delay in the dissemination of new information and research findings in the field of spoken language processing. We are, of course, very interested in following the work of other colleagues who are carrying out research on speech perception and spoken language processing and we would be grateful if you and your colleagues would send us copies of any recent reprints, preprints and progress reports as they become available so that we can keep up with your latest findings. Please address all correspondence to:

Professor David B. Pisoni
Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405-1301
United States of America

Telephone: (812) 855-1155, 855-1768
Facsimile: (812)855-4691
E-mail: pisoni@indiana.edu
Web: http://www.indiana.edu/~srlweb

Copies of this progress report are being sent primarily to libraries and specific research institutions rather than individual scientists. Because of the rising costs of publication and printing, it is not possible to provide multiple copies of this report to people at the same institution or issue copies to individuals. We are eager to enter into exchange agreements with other institutions for their reports and publications. Please write to the above address for further information.

The information contained in this progress report is freely available to the public and is not restricted in any way. The views expressed in these research reports are those of the individual authors and do not reflect the opinions of the granting agencies or sponsors of the specific research.

Speech Research Laboratory Faculty, Staff & Technical Personnel

(January 1, 1996–December 31, 1997)

Research Personnel

David B. Pisoni, Ph.D.	Chancellors' Professor of Psychology and Cognitive Science[1,2]
Karen Iler Kirk, Ph.D.	Assistant Professor of Otolaryngology-Head & Neck Surgery[3]
Mario Svirsky, Ph.D.	Associate Professor of Otolaryngology-Head & Neck Surgery[3]
Ann R. Bradlow, Ph.D.	Assistant Research Scientist[4]
Steven B. Chin, Ph.D.	Assistant Scientist in Otolaryngology-Head & Neck Surgery[3]
Brian F. Bowdle, Ph.D.	COAS Post-doctoral Trainee
Stefan A. Frisch, Ph.D.	NIH Post-doctoral Trainee
Ted A. Meyer, M.D., Ph.D.	NIH Post-doctoral Trainee[3]
Sonya M. Sheffert, Ph.D.	NIH Post-doctoral Trainee
Michael S. Vitevitch, Ph.D.	NIH Post-doctoral Trainee
Richard A. Wright, Ph.D.	NIH Post-doctoral Trainee
Nathan E. Amos, M.S.	NIH Pre-doctoral Trainee
Miranda Cleary, A.B.	NIH Pre-doctoral Trainee
Winston D. Goh, M.Soc.Sci.	Pre-doctoral Trainee[5]
Lorin Lachs, B.A.	NIH Pre-doctoral Trainee
Michele L. Morrisette, B.S.	NIH Pre-doctoral Trainee
Martin Rickert, M.Sc.	NIH Pre-doctoral Trainee
Carolyn L. Pytte, B.S.	NIH Pre-doctoral Trainee
Jill E. Beavins, B.S.	NIH Medical Student Trainee
Matthew D. Caldwell, B.S.	NIH Medical Student Trainee
Jason L. Niksch, B.A.	NIH Medical Student Trainee
Steven Rider, B.A.	NIH Medical Student Trainee
Robert B. Sloan, M.S.	NIH Medical Student Trainee
Gary A. Wright, B.S.	NIH Medical Student Trainee

---

[1] Also Adjunct Professor of Linguistics, Indiana University, Bloomington, IN

[2] Also Adjunct Professor of Otolaryngology-Head & Neck Surgery, Indiana University School of Medicine, Indianapolis, IN

[3] Department of Otolaryngology-Head & Neck Surgery, Indiana University School of Medicine, Indianapolis, IN

[4] Department of Communication Sciences and Disorders, Northwestern University, Evanston, IL

[5] Also Senior Tutor, Department of Social Work and Psychology, National University of Singapore

Andrew H. Bangert	Undergraduate Research Assistant
Janna L. Carlson	Undergraduate Research Assistant
Erin E. Colone	Undergraduate Research Assistant
James J. Brink	Undergraduate Research Assistant
Melissa Kluck	Undergraduate Research Assistant
Nathan R. Large	Undergraduate Research Assistant

Technical Personnel

Luis R. Hernández, B.A.	Research Associate in Psychology/Systems Administrator
Darla J. Sallee	Administrative Assistant
Jerry C. Forshee, M.A.	Computer Systems Analyst[6]
David A. Link	Electronics Engineer
Gina M. Torretta, B.A.	Research Technician
Christopher D. Quillet, B.S.	Research Technician
Monica Wright, B.A.	Graphics Technician
Jon M. D'Haenens	Programmer
Vargha S. Manshadi	Programmer
Caleb Marcinkovich	Programmer

E-Mail Addresses

pisoni@indiana.edu

kkirk@iupui.edu
msvirsky@iupui.edu

abradlow@indiana.edu
schin@iupui.edu

bbowdle@indiana.edu
ssheffer@indiana.edu
safrisch@indiana.edu
mvitevit@indiana.edu
tmeyer@iupui.edu
riawrigh@indiana.edu

namos@indiana.edu
mmorrise@indiana.edu
micleary@indiana.edu
cpytte@indiana.edu
llachs@indiana.edu
wigoh@indiana.edu

abangert@indiana.edu
ecolone@indiana.edu
jbrink@indiana.edu
nlarge@indiana.edu
jcarlso@indiana.edu

hernande@indiana.edu
cquillet@indiana.edu
vmanshad@indiana.edu
dsallee@indiana.edu
mwright@psysrl.psych.indiana.edu
forshee@indiana.edu

---

[6] Also the Director of Technical Support, Department of Psychology, Indiana University, Bloomington, IN

I. Extended Manuscripts

Speech Perception[1]

Richard Wright, Stefan Frisch, and David B. Pisoni

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405

---

[1] This work supported by NIH/NIDCD Training Grant DC00012 to Indiana University.

Speech Perception

Introduction

The study of speech perception is concerned with the process by which the human listener, as a participant in a communicative act, derives meaning from spoken utturances. Modern speech research began in the late 1940s, and the problems that researchers in speech perception have focused on have remained relatively unchanged since. They are: 1) variability in the physical signal and the search for acoustic invariants, 2) human perceptual constancy in the face of diverse physical stimulation, and 3) the neural representation of the speech signal. The goal of this chapter is to examine how these problems have been addressed by various theories of speech perception and describe how basic assumptions about the nature of the problem have shaped the course of research. Due to the breadth of information to be covered, this chapter will not examine the specifics of experimental methodology or survey the empirical literature in the field. There have been many detailed reviews of speech perception which supply further background on these topics (e.g., Studdert-Kennedy, 1974, 1976; Darwin, 1976; Pisoni, 1978; Lively, Pisoni, & Goldinger, 1994; Klatt, 1989; Miller, 1990; Goldinger, Pisoni, & Luce, 1996; Neary, 1997; Nygaard & Pisoni 1995).

The process of speech perception may be limited to the auditory channel alone as in the case of a telephone conversation. However, in everyday spoken language the visual channel is also involved as well and the study of multi-modal speech perception and spoken language processing is one of the central areas of current research. While stimulus variability, perceptual constancy, and neural representation are core problems in all areas of perception research, speech perception is unlike other perceptual processes because the perceiver also produces spoken language and therefore has intimate knowledge of the signal source. This relationship, combined with the high communicative load of speech constrains the signal significantly and affects both perception and production strategies (Lieberman 1963; Fowler & Housman, 1987; Lindblom, 1990). Speech perception is also unique in its remarkable robustness in the face of a wide range of environmental and communicative conditions. The listener's remains remarkably constant in the face of a significant amount of production related variation in the signal. Furthermore, even in the worst of environmental conditions in which large portions of the signal are distorted or masked, the spoken message is recovered with little or no error. As we shall see, part of this perceptual robustness derives from the richness and redundancy of information in the signal, part of it lies in the highly structured nature of language, and part comes from the context dependent nature of spoken language.

Extracting meaning from the acoustic signal may at first glance seem like a relatively straightforward task. It would seem to be simply a matter of identifying the acoustically invariant characteristics in the frequency and time domains of the signal that correspond to the appropriate serially ordered linguistic units (i.e. reversing the encoding of those mental units by the production process). From those units the hearer can then retrieve the appropriate lexical entries from memory. Although stated rather simply here, this approach is based on an assumption about the process of speech perception that has been at the core of most symbolic processing approaches (Studdert-Kennedy, 1976). That is, the process involves the segmentation of the signal into discrete and abstract linguistic units such as features, phonemes, or syllables. Before or during segmentation the extra-linguistic information is segregated from the intended message and is processed separately or discarded. For this process to succeed, the spoken signal must meet two conditions The first, known as the *invariance condition*, is that there is invariant information in the signal that is present in all instances that correspond to the perceived linguistic unit. The second, known as the *linearity condition*, is that the information in the signal is serially ordered so that information about the first linguistic unit precedes and does not completely overlap or follow information about the next linguistic unit and so forth.

It has become apparent to speech researchers over the last 40 years that the invariance and linearity conditions are almost never met in the actual speech signal (Liberman, 1957; Chomsky & Miller, 1963; Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967). This has led to several innovations that have achieved varying degrees of success in accommodating some of the variability and much of the nonlinearity inherent in the speech signal (Liberman, Cooper, Harris, & MacNeilage, 1963; Liberman & Mattingly, 1985; Blumstein & Stevens, 1980; Stevens & Blumstein, 1981). However, inter- and intra-talker variability remains an intractable problem within these conceptual/theoretical frameworks. Recent approaches that treat the signal holistically have proven promising alternatives. Much of the variability that researchers sought to strip away in traditional approaches contains important information about the talker and about the intended message. Recent approaches, while differing significantly in their view of perception, treat the signal as information rich. The information in the speech signal is both 'linguistic', the traditional message of the signal, and 'non-linguistic' or 'indexical' (Abercrombie, 1967; Ladefoged & Broadbent, 1957), information about the talker's immediate physical and emotional state, about the talker's relationship to the environment, the social context, etc. (Pisoni, 1996). Much of the variability and redundancy in the signal can be used to enhance the perceptual process rather than being discarded as noise (Klatt, 1976, 1989; Fowler, 1986; Goldinger, 1990; Johnson, 1997).

The Abstractionist/Symbolic Approach to Speech Perception

Traditional approaches to speech perception are based on ideas that originated in information theory and have treated the process of speech perception as distinct from word recognition, sentence understanding, and speaker recognition. In this view, the decoding of the speech signal into abstract symbolic units (i.e. features, phonemes, syllables) is the goal of speech perception, and the discrete units are then passed along to be used by higher level parsers that identify lexical items such as morphemes or words. Listeners are hypothesized to extract abstract, invariant properties of the acoustic signal to be matched to prototypical representations stored in long term memory (Forster, 1976; McClelland & Elman, 1986; Oden & Massaro, 1978; Samuel, 1982; Kuhl, 1991; Neary, 1992). In fact, most models of word recognition use either the phoneme or the syllable as the fundamental unit of processing (Marslen-Wilson & Welsh, 1978; Cutler & Norris, 1988; Norris & Cutler, 1995; Luce, 1986). These models implicitly assume some type of low-level recoding process.

The assumption that speech is perceived in abstract idealized units has led researchers to search for simple first-order physical invariants and to ignore the problem of stimulus variability in the listener's environment (e.g., Blumstein & Stevens, 1980; Sussman, McCaffrey, & Matthews, 1991). In this view, variability is treated as noise. This means that much of the talker specific characteristics, or *indexical* information, that a listener uses to identify a particular talker, or a talker's state, is removed through a process of normalization, leaving behind the intended linguistic message (Studdert-Kennedy, 1974). In this view, normalization converts the physical signal to a set of abstract units that represent the linguistic message symbolically.

The dissociation of form from content in speech perception has persisted in large part despite the fact that the both sources of information are carried simultaneously and in parallel in the acoustic signal and despite the potential gain that a listener may get from simultaneously receiving contextual information such as the rate of an utterance, or the gender, socioeconomic status, and mood of the talker. Following models of concept learning and memory (Jacoby & Brooks, 1984), this view of speech perception has been termed the *abstractionist* approach (Nygaard & Pisoni, 1995; Pisoni, 1997). As the abstractionist approach relies on a set of idealized linguistic units, it is useful to review the types of perceptual units that are commonly used and the motivations for abstract units in the first place.

The use of abstract symbolic units in almost all traditional models of speech perception came about for several reasons, one being that linguistic theory has had a great impact on speech research. The abstract units that had been proposed as tools for describing patterns of spoken language, themselves a reflection of the influence of information theory on linguistics (Jacobson, Fant, & Halle, 1952), were adopted by many speech researchers (e.g., Liberman et al., 1957; Pisoni, 1978). This view can be summed up by a quote from Halle (1985):

> when we learn a new word we practically never remember most of the salient acoustic properties that must have been present when the acoustic signal struck our ears; for example, we do not remember the voice quality, the speed of utterance, and other properties directly linked to the unique circumstances directly surrounding every utterance. (p. 101)

While linguistic theory has moved away from the phoneme as a unit of linguistic description to a temporally distributed featural or gestural array (Browman & Goldstein, 1990; Goldsmith, 1990; Steriade, 1993; Fowler, 1995; Frazier, 1995), many researchers in speech perception continue to use the phoneme as a unit of perception.

Another reason for the use of abstract units lies in the nature of the speech signal. Because of the way speech is produced in the vocal tract, the resulting acoustic signal is continuously changing, making all information in the signal highly variable and transient. This variability, combined with the constraints on auditory memory, led many researchers to assume that the analog signal must be rapidly recoded into discrete and progressively more abstract units (Broadbent, 1965; Liberman et al., 1967). This process achieves a large reduction of data that was thought to be redundant or extraneous into a few predefined and timeless dimensions. However, while reduction of redundancy potentially reduces the memory load, it increases the processing load and greatly increases the potential for an unrecoverable error on the part of the hearer (Miller, 1962; Klatt, 1979). Furthermore, there is evidence that much of the information in the signal that was deemed extraneous is encoded and stored by the memory system and subsequently used by the hearer in extracting meaning from the spoken signal (Peters, 1955; Creelman, 1957; Pisoni 1990; Palmeri, Goldinger & Pisoni, 1993).

An additional motivation for postulating abstract units comes from the phenomenon of perceptual constancy. Although there is substantial contextual variation in the acoustic signal, the hearer appears to perceive a single unit of sound. For example, a voiceless stop consonant such as /t/ that is at the beginning of a word, as in the word "top", is accompanied by a brief puff of air at its release and a period of voicelessness in the following vowel which together are generally referred to as *aspiration*. When that same stop is preceded by the fricative /s/, as in the word "stop", the aspiration is largely absent. Yet the hearer perceives the two very different acoustic signals as being the same sound category /t/. This particular example of perceptual constancy may be explained in terms of the possible lexical contrasts of English. While there are lexical distinctions that are based on the voicing contrast, "cat" versus "cad" for example, there is no lexical distinction in English that is based on an aspiration contrast. It should be remembered that most contextual variation that is non-contrastive in one language is often the foundation of a lexical contrast in another language (Ladefoged & Maddieson, 1996). Rather than being hardwired into the brain at birth or being imposed on the hearer by transformations of the peripheral auditory system, these contrastive characteristics of a particular language must be learned. Thus, the complex process by which perceptual normalization takes place in a particular language is due almost entirely to perceptual learning and categorization.

Finally, segmenting the speech signal into units that are hierarchically organized permits a duality of patterning of sound and meaning (Hockett, 1960) that is thought to give language its communicative power. That is, smaller units such as phonemes may be combined according to language specific phonotactic constraints into morphemes and words, and words may be organized according to grammatical constraints into sentences. This means that with a small set of canonical sound units, and the possibility of recursiveness, the talker may produce and the hearer may decode and parse a virtually unbounded number of utterances in the language. There are many types of proposed abstract linguistic units that are related in a nested structure with features at the terminal nodes and other types of units that may include phonemes, syllables, morphemes, words, syntactic phrases and intonation phrases as branching nodes that dominate them (Jacobson, Fant & Halle, 1952; Chomsky & Halle, 1968; Pierrehumbert & Beckman, 1988).

Different approaches to speech perception employ different units and different assumptions about levels of processing. Yet, there is no evidence for the primacy of any particular unit in perception. In fact, the perceptual task itself may determine the units that hearers use to analyze the speech signal (Miller, 1962; Cutler, 1997). There are many behavioral studies that have found that human listeners appear to segment the signal into phoneme sized units. For example, it has been found that reaction times of English listeners to phonotactically permissible CVC syllables (where all the sounds in isolation are permissible) were no faster than reaction times to phonotactically impermissible CV syllables indicating that the syllable plays no role in spoken language processing (Cutler, Mehler, Norris, & Segui, 1986; Norris & Cutler, 1988). However, the same experiments conducted with native speakers of French have found that listeners' response times are significantly more rapid to the phonotactically permissible CVC syllable than to the CV syllables, while responses of Japanese listeners to the CVC were significantly slower than to the CV syllables (Mehler, 1981; Mehler, Dommergues, Frauenfelder and Segui, 1981; Segui, 1984). Taken together findings of task specific and language specific biases in the preferred units of segmentation indicate that a particular processing unit is contingent on a number of factors. Moreover, there is a great deal of evidence that smaller units like the phoneme or syllable are perceptually contingent on larger units such as the word or phrase (Miller, 1962; Bever, Lackner & Kirk, 1969; Ganong, 1980). This interdependence argues against the strict hierarchical view of speech perception in which the smallest units are extracted as a precursor to the next higher level of processing (Remez, 1987). Rather, the listeners' responses appear to be sensitive to attentional demands, processing contingencies, and available information (Eimas & Nygaard, 1992; Remez, Rubin, Berns, Pardo & Lang, 1994).

Basic Stimulus Properties

Understanding the nature of the stimulus is an important step in approaching the basic problems in speech perception. This section will review some of the crucial findings that are relevant to models of speech perception. A large portion of the research on speech perception has been devoted to the investigation of speech cues and some of the better known findings are discussed in four sub-sections: vowels, consonant place, consonant manner, and consonant voicing. In addition to the auditory channel, the visual channel is known to affect speech perception, and we discuss below some of the key findings. Because the speech signal is produced by largely overlapping articulatory gestures, information in the signal is distributed in an overlapping fashion. Non-linearity of information in the speech signal is reviewed and implications for speech perception are touched upon. Finally, although it was largely ignored in the past, variability is arguably the most important issue in speech perception research. It is a problem comes from many sources; some of the most important sources of variability in speech and their perceptual consequences are reviewed in the last part of this section.

Speech Cues

Since the advent of modern speech research at the end of the Second World War, much of the work on speech perception has focused on identifying aspects of the speech signal which contain the minimal information that is necessary to convey a speech contrast. These components of the signal are referred to as *speech cues*. The assumption that a small set of acoustic features or attributes in the acoustic signal provide cues to linguistic contrasts was the motivation for the search for invariant cues which was central to speech perception research from the mid-50's up until the present time. The more the signal was explored for invariant acoustic cues the more the problems of variability and non-linearity became evident. One solution to this problem was to study speech using highly controlled stimuli to minimize the variability. Stimuli for perception experiments were usually constructed using a single synthetic voice producing words in isolation with a single contrast in one segmental context and syllable position. This approach produced empirical evidence about a very limited set of circumstances, thereby missing much of the systematic variability and redundancy of information that plays a part in speech perception. It also artificially removed talker specific information and other extra-linguistic contextual information that was later shown to be used by listeners during speech perception. Despite these shortcomings, early work on speech perception has provided valuable empirical data that must be taken into consideration when evaluating the relative merits of current speech perception models.

The acoustic signal is produced by articulatory gestures that are continuous and overlapping to various degrees; thus, the resulting acoustic cues vary greatly with context, speaking rate, and talker. Contextual variation is a factor that contributes to redundancy in the signal. Although the early study of speech cues sought to identify a single primary cue to a particular linguistic contrast, it is improbable that the human perceptual system fails to take advantage of the redundancy of information in the signal. It should also be noted that many factors may contribute to the salience of a particular acoustic cue in the identification of words, syllables, or speech sounds. These include factors that are part of the signal such as coarticulation or positional allophony, semantic factors such as the predictability of a word that the listener is trying to recover, and whether or not the listener has access to visual information generated by the talker. The extent to which a listener attends to particular information in the speech signal is also dependent on the particular system of sound contrasts in his/her language. For example, while speakers of English can use a lateral/rhotic contrast (primarily in F3) to distinguish words like "light" from words like "right", many languages lack this subtle contrast. Speakers of those languages (e.g., Japanese) have great difficulty attending to the relevant speech cues when trying to distinguish /l/ from /r/. Thus, the speech cues that are discussed in this section should not be considered invariant or universal, instead they are context sensitive and highly interactive.

Vocalic Contrasts

The vocal tract acts as a time-varying filter with resonant properties that transform frequency spectra of the sound sources generated in the vocal tract (Fant, 1960; Flanagan, 1972; Stevens & House, 1955). Movements of the tongue, lips, and jaw cause changes in the resonating characteristics of the vocal tract that are important in distinguishing one vowel from another. Vowels distinctions are generally thought to be based in part on the relative spacing of the fundamental frequency (f0) and the first three vocal tract resonances or formants (F1, F2, F3) (Syrdal & Gopal, 1986). In general, there is an inverse relationship between the degree of constriction (vowel height) and the height of the first formant (Fant, 1960). That is, as the degree of constriction in the vocal tract increases, increasing the vowel height, F1 lowers in frequency. The second formant is generally correlated with the backness of the vowel: the further back in the vocal tract the vowel constriction is made the lower the second formant (F2). F2 is also lowered by lip rounding and protrusion. Thus, the formant frequencies of vowels and vowel-like sounds are produced by changes in the length and shape of the resonant cavities of the vocal tract above the laryngeal sound source.

In very clear speech vowels contain steady-state portions where the relative spacing between the formants remains fixed and the f0 remains relatively constant. Words spoken with care and speech samples from read sentences may contain steady-state vowels. Early work on vowel perception that was modeled on this form of carefully articulated speech found the primary cue for vowel perception was the steady-state formant values (Gerstman, 1968; Skinner, 1977). Figure 1 is a spectrogram illustrating the formant structure of five carefully pronounced nonsense words with representative vowels dVd contexts: /did/ (sounds like "deed"), /ded/ (sounds like "dayed"), /dad/ (sounds like "dodd"), /dod/ (sounds like "doad"), and /dud/ (sounds like "dood"). The first two formants, the lowest two dark bars, are clearly present and have relatively steady state portions near the center of each word. The words /did/ and /dad/ have the clearest steady state portions.

Insert Figure 1 about here

In naturally spoken language, however, formants rarely achieve a steady-state, and are usually flanked by other speech sounds which shape the formant structure into a dynamic time varying pattern. For the same reasons, vowels often fall short of the formant values observed in careful speech resulting in 'undershoot' (Fant, 1960; Stevens & House, 1963). The degree of undershoot is a complex function of the flanking articulations, speaking rate, prosody, sentence structure, dialect, and individual speaking style (Lindblom, 1963; Gay, 1978). These observations have led researchers to question the assumption that vowel perception relies on the perception of steady-state formant relationships, and subsequent experimentation has revealed that dynamic spectral information in the formant transitions into and out of the vowel are sufficient to identify vowels even in the absence of any steady-state information (Lindblom & Studdert-Kennedy, 1967; Strange, Jenkins, & Johnson, 1983).

Although they are not found in English, and not well studied in the perception literature, there are many secondary vocalic contrasts in the world's languages in addition to vowel height and backness contrasts. A few of the more common ones are briefly described here; For a complete review see Ladefoged & Maddieson (1996). A secondary contrast may effectively double or, in concert with other secondary contrasts, triple or quadruple the vowel inventory of a language. The most common type of secondary contrast, found in 20% of the world's languages, is nasalization (Maddieson, 1984). Nasalization in speech is marked by a partial attenuation of energy in the higher frequencies, a broadening of formant bandwidths and by an additional weak nasal formant around 300 Hz (Fant 1960; Fujimura, 1962). Vowel length contrasts are also commonly observed and are found in such diverse languages as Estonian (Lehiste, 1970), Thai (Hudak, 1987) and Japanese (McCawley, 1968). *Source characteristics*, changes in the vibration characteristics of the vocal folds, may also serve as secondary contrasts. These include *creaky* and *breathy* vowels. In a creaky vowel, the vocal fold vibration is characterized by a smaller open-to-closed ratio resulting in more energy in the harmonics of the first and second formants, narrower formant bandwidths, and often more *jitter* (irregular vocal cord pulse rate) (Ladefoged, Maddieson, & Jackson, 1988). In breathy vowels, the vocal fold vibration is characterized by a greater open-to-closed ratio resulting in more energy in the fundamental frequency, broader formant bandwidths, and often more random energy (noise component) (Ladefoged et al, 1988). Although they are poorly understood, secondary contrasts appear to interact in complex ways with perceived vowel formant structure and with the perception of voice pitch (Silverman, 1997).

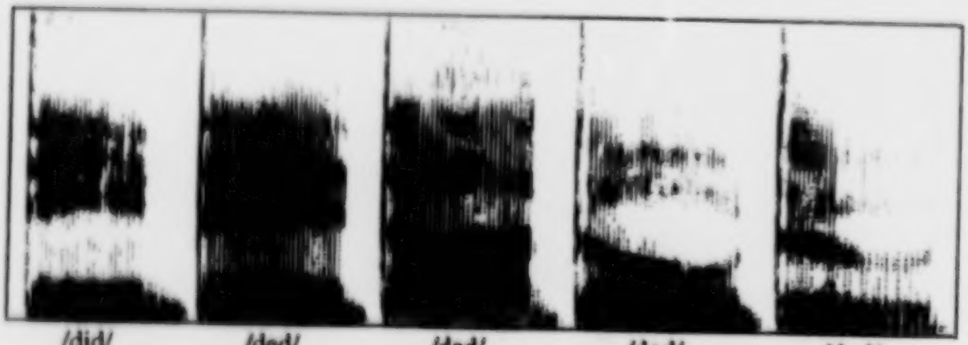


Figure 1. A spectrogram illustrating the formant structure of four representative vowels in dVd contexts. The lowest two dark bands are the first and second formants.

18

Consonant Place of Articulation

There are several potential sources of cues to the place of articulation of a consonant, including second formant transitions, stop release bursts, nasal pole-zero patterns, and the generation of fricative noise. The strongest place of articulation cues are found in the brief transitional period between a consonant and an adjacent vowel. Some speech cues are internal, as is the case in fricatives such as /s/ or nasals such as /n/. Other speech cues are distributed over an entire syllable as in vowel 'coloring' by laterals such as /l/, rhotics such as /r/, and retroflex consonants such those found in Malayam and Hindi. We briefly review several of the most important cues to place of articulation here.

Formant Transitions

The second formant (F2), and to a lesser degree the third formant (F3), provide the listener with perceptual cues to the place of articulation of consonants with oral constrictions, particularly the stops, affricates, nasals, and fricatives (Delattre, Liberman & Cooper, 1955). Transitions are the deformation of the vowel's formants resulting from the closure or aperture phase of a consonant's articulation, i.e., a rapid change in the resonating cavity, overlapping with the relatively open articulation of a flanking vowel. Because they are the result of very fast movements of the articulators from one position to another, formant transitions are transient and dynamic, with the speed of the transitions depending on the manner, the place of articulation (to a lesser degree), and such factors as the individual talker's motor coordination, the speaking rate and style, and the novelty of the utterance.

Unlike other consonants, glides and liquids have clear formant structure throughout their durations. Glides are distinguished from each other by the distance between the first and second formant values at the peak of constriction, whereas the English /l/ is distinguished from /r/ by the relative frequency of the third formant (O'Connor, Gerstman, Liberman, Delattre, & Cooper, 1957). English /l/ and /r/ cause *vowel coloring*, a change in the formant structure of the adjacent vowels, particularly the preceding one, that may last for much of the vowel's duration.

Both the transitions into and out of the period of consonant constriction provide place cues for consonants that are between vowels. In other positions, there is at most only a single set of formant transitions to cue place: the formant transitions out of the consonant constriction (C to V) in word onset and post-consonantal positions, and the formant transitions into the consonant constriction (V to C) in word final and pre-consonantal positions. For stops in the VC position with no *audible* release, formant transitions may provide the only place cues. Following the release of voiceless stops, there is a brief period of voicelessness during which energy in the formants is weakened. Following the release of aspirated stops, a longer portion or all of the transition may be present in a much weaker form in the aspiration noise. It is widely thought that CV formant transitions provide more salient information about place than VC transitions (see Wright, 1996, Ch. 2 for a discussion). When formant transitions into a stop closure (V to C) conflict with the transitions out of the closure (C to V), listeners identify the stop as having the place of articulation that corresponds with the C to V transitions (Fujimura, Macchi, & Streeter, 1978). The relative prominence of CV transitions over VC transitions is also influenced by the language of the listener. For example, native Japanese speaking listeners have been shown to be very poor at distinguishing place from VC transitions alone, while native Dutch and English speakers are good at distinguishing place with VC transitions (van Weiringen, 1995). In this case, the difference in performance can be attributed to differences in syllable structure between the languages. While English and Dutch allow post-vocalic stops with contrasting place of articulation (e.g., "actor" or "bad"), Japanese does not; experience with Japanese syllable structure has biased Japanese speakers towards relying more on the CV transitions than VC transitions.

Fricative Noise

Fricatives are characterized by a narrow constriction in the vocal tract that results in turbulent noise either at the place of the constriction or at an obstruction down-stream from the constriction (Schadle, 1985). Frication noise is aperiodic with a relatively long duration. Its spectrum is shaped primarily by the cavity in front of the noise source (Heinz and Stevens, 1961). The spectrum of the frication noise is sufficient for listeners to reliably recover the place of articulation in sibilant fricatives such as /s/ and /z/. However, in other fricatives with lower amplitude and more diffuse spectra, such as /f/ and /v/, the F2 transition has been found to be necessary for listeners to reliably distinguish place of articulation (Harris, 1958). Of these, the voiced fricatives, as in the words "that" and "vat" are the least reliably distinguished (Miller & Nicely, 1955). It should be noted that this labio-dental versus inter-dental contrast in fricatives in English is very rare in the world's languages (Maddieson, 1984). The intensity of frication noise and the degree of front cavity shaping is expected to affect the relative importance of the fricative noise as a source of information for other fricatives as well.

Because fricatives have continuous noise that is shaped by the cavity in front of the constriction, they also convey information about adjacent consonants in a fashion that is similar to vowels. Overlap with other consonant constrictions results in changes in the spectral shape of a portion of the frication noise, most markedly when the constriction is in front of the noise source. The offset frequency of the fricative spectrum in fricative-stop clusters serves as a cue to place of articulation of the stop (Bailey & Summerfield, 1980; Repp & Mann, 1981).

Stop Release Bursts

In oral stop articulations there is complete occlusion of the vocal tract and a resulting build-up of pressure behind the closure. The sudden movement away from complete stricture results in brief high amplitude noise known as the *release burst* or *release transient*. Release bursts are aperiodic with a duration of approximately 5-10 ms. The burst's duration depends on both the place of articulation of the stop and the quality of the following vowel; Velar stop releases (/k/ and /g/) are longer and noisier than labial and dental stops, and both dental and velar stops show an increased noisiness and duration of release before high vowels. Release bursts have been shown to play an important role in the perception of place of articulation of stop consonants (e.g., Liberman et al., 1955, Blumstein, 1981; Dorman, Studdert-Kennedy, & Raphael, 1977; Kewley-Port, 1983a). Although the release burst or the formant transitions alone are sufficient cues to place, the formant transitions have been shown to dominate place perception; i.e. if the release burst spectrum and the F2 transition provide conflicting place cues, listeners perceive place according to the F2 transition (Walley & Carrell, 1983). Listeners show the greatest reliance on the transition in identifying velar place in stops (Kewley-Port, Pisoni, & Studdert-Kennedy, 1983). Although less studied as a source of cues, there are many other subtler place dependent differences among stops that are a potential source of information to the listener. For example, velar stops (/k/ and /g/) tend to have shorter closure durations than labial stops (/p/ and /b/) and amplitude differences may help in distinguishing among fricatives.

An additional class of sounds known as *affricates* are similar in some respects to stops and in other aspects to fricatives; they have a stop portion followed by a release into a fricative portion. In their stop portion, they have a complete closure, a build up of pressure and the resultant release burst at release. The release is followed by a period of frication longer than stop aspiration but shorter than a full fricative. Both the burst and the frication provide place cues. In English, all affricates are palato-alveolar, but there is a voicing contrast ("chug" versus "jug"). The palato-alveolar affricate found in English is the most commonly found, 45 percent of the world's languages have it (Maddieson, 1984), but many other places of articulation are common and many languages have an affricate place contrast (e.g., /pf/ versus /ts/ in German).

Nasal Cues

Like the oral stops, nasal consonants have an oral constriction that results in formant transitions in the adjacent vowels. In addition, nasals show a marked weakening in the upper formants due to the antiresonance (zero) and a low frequency resonance (pole) below 500 Hz. The nasal pole-zero pattern serves as a place cue (Kurowski & Blumstein, 1984). This cue is most reliable in distinguishing /n/ and /m/, and less so for other nasals (House, 1957). Listeners identify the place of articulation more reliably from external formant transitions than from the internal nasal portion of the signal (Malécot, 1956). Figure 2 schematically illustrates some of the most frequently cited cues to consonant place of articulation for three types of consonants: a voiceless alveolar stop /t/, a voiceless alveolar fricative /s/, and an alveolar nasal /n/. The horizontal bars represent the first three formants of the vowel /a/, the deformation of the bars represents formant transitions, and the hatched areas represent fricative and stop release noises. Table 1 summarizes the consonant place cues discussed above. Although it lists some of the more frequently discussed cues, the table should not be seen as exhaustive as there are many secondary and contextual cues that contribute to a consonant percept that are not listed here.

Insert Figure 2 about here

Table 1

Summary of Commonly Cited Place Cues

Cue	**Applies to**	**Distribution**
F2 transition	all	VC, CV transitions
burst spectrum	stops	C-release
frication spectrum	fricatives, affricates (esp. sibilants)	internal
frication amplitude	fricatives	internal
nasal pole, zero	nasals	internal
fricative noise transition	stops	fricative edge
F3 height	liquids and glides	internal

Consonants of all types have much narrower constrictions than vowels. They can be viewed as the layering of a series of rapid constricting movements onto a series of slower moving transitions from one vowel to the next (Browman & Goldstein, 1990). For all types of consonants, the change in the vowel's formants that result from the influence of the consonant's narrower constriction are the most robust types of cues. However, we have seen that there are a number of other sources of information about the place of articulation of a consonant that the listener may use in identifying consonants. This chapter has touched on a few of the better known such as stop release bursts, nasal pole-zero patterns, and fricative noise. These are often referred to as 'secondary' cues because perceptual tests have shown that when paired with formant transitions that provide conflicting information about the consonant place of articulation, the

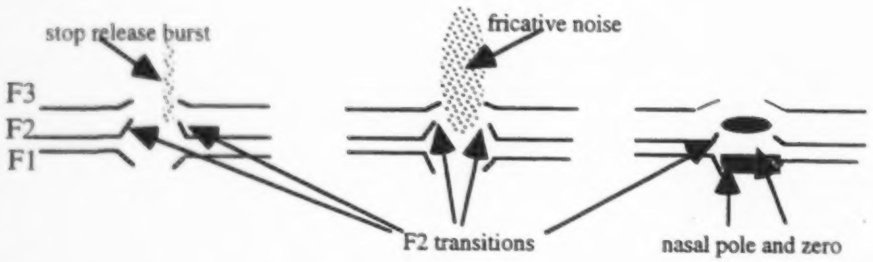

stop release burst
fricative noise
F3
F2
F1
F2 transitions
nasal pole and zero

Figure 2. Schematic illustration of place cues in three VCV sequences where V is the vowel /a/ and C is a voiceless alveolar stop, a voiceless alveolar fricative and an alveolar nasal

perceived place is that appropriate for the formant transitions. However, depending on the listening conditions, the linguistic context, and the perceptual task, these so-called secondary cues may serve as the primary source of information about a consonant's place of articulation. For example, in word initial fricative-stop clusters, the fricative noise may provide the sole source of information about the fricative's place of articulation. While in English only /s/ appears in word initial fricative-stop clusters, there are many languages which contrast fricative place in such clusters and many which have stop-stop clusters or nasal-stop clusters as well (see Wright, 1996 for a description of a language that has all three types of clusters). Thus, it is likely that phonotactic constraints, position within sentence, position within word, position within syllable, background noise, and so on, must be taken into consideration before a relative prominence or salience is assigned to any particular acoustic cue in the signal.

Consonant Manner Contrasts

All oral constrictions result in an attenuation of the signal, particularly in the higher frequencies. The relative degree of attenuation is a strong cue to the manner of a consonant. An abrupt attenuation of the signal in all frequencies is a cue to the presence of a stop. Insertion of a period of silence in a signal, either between vowels or between a fricative and a vowel can result in the listener perceiving a stop (Bailey & Summerfield, 1980). A complete attenuation of the harmonic signal together with fricative noise provides the listener with cues to the presence of a fricative. A less severe drop in amplitude accompanied by nasal murmur and a nasal pole and zero are cues to the presence of a nasal (Hawkins & Stevens, 1985). Nasalization of the preceding vowel provides 'look-ahead' cues to post-vocalic nasal consonants (Ali, Gallager, Goldstein, & Daniloff, 1971; Hawkins & Stevens, 1985).

Glides and liquids maintain formant structure throughout their peak of stricture, but both attenuate the signal more than vowels. Glides are additionally differentiated from other consonants by the relative gradualness of the transitions into and out of the peak of stricture. Lengthening the duration of synthesized formant transitions has been shown to change the listener's percept of manner from stop to glide (Liberman, Delattre, Gerstman, & Cooper, 1956). A similar cue is found in the amplitude envelope at the point of transition between consonant and vowel: stops have the most abrupt and glides have the most gradual amplitude rise time (Shinn & Blumstein, 1984).

Manner cues in general tend to be more robust than place cues because they result in more salient changes in the signal, although distinguishing stop from fricative manner is less reliable with the weaker fricatives (Miller & Nicely, 1955). Figure 3 schematically illustrates some of the most frequently cited cues to consonant manner of articulation for three types of consonants: a voiceless alveolar stop /t/, a voiceless alveolar fricative /s/, and an alveolar nasal /n/. The horizontal bars represent the first three formants of the vowel /a/, the deformation of the bars represents formant transitions, the hatched areas represents fricative and stop release noises. Table 2 summarizes the consonant manner cues discussed above. Again, the table should not be seen as exhaustive as there are many secondary and contextual cues that contribute to a consonant percept that are not listed here.

Insert Figure 3 about here

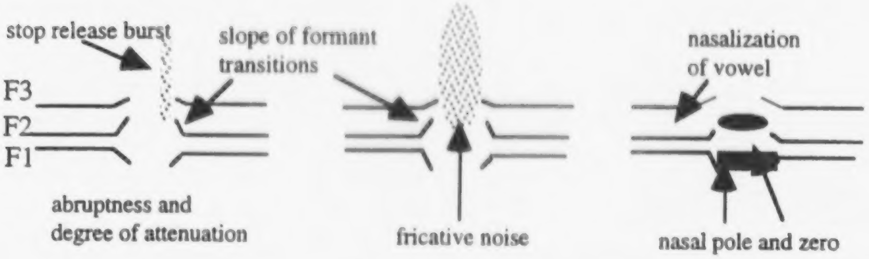

stop release burst
slope of formant transitions
nasalization of vowel
F3
F2
F1
abruptness and degree of attenuation
fricative noise
nasal pole and zero

Figure 3. Schematic illustration of manner cues in three VCV sequence where V is the vowel /a/ and C is a voiceless alveolar stop, a voiceless alveolar fricative and an alveolar nasal

Table 2

Summary of Commonly Cited Manner Cues

Cue	Applies to	Distribution
silence/near silence	stops, affricates	internal
frication noise	fricatives, affricates	internal
nasal pole & zero	nasals	internal
vowel nasalization	nasals	adjacent vowel
formant structure	liquids, glides (vowels)	internal
release burst	stops	C-release
noise duration	stop, affricate, fricative	internal
noise onset rise-time	stop/affricate, fricative	internal
transition duration	stop, glide	VC, CV transitions

Cues to Voicing Contrasts

Vocal fold vibration, resulting in periodicity in the signal, is the primary cue to voicing; however tight oral constriction inhibits the airflow necessary for vocal fold vibration. In English and many other languages, voiced obstruents, especially stops, may have little or no vocal fold activity. This is more common for stops in syllable final position. In this situation, the listener must rely on other cues to voicing. There are several other important cues such as Voicing Onset Time (VOT), the presence and the amplitude of aspiration noise, and durational cues. For syllable initial stops in word onset position, the primary cue appears to be VOT. This is not really a single cue in the traditional sense but a dynamic complex that includes the time between the release burst and the onset of vocal fold vibration together with aspiration noise, i.e. low amplitude noise with spectral peaks in the regions of the following vowel's formants (Lisker & Abramson, 1964). VOT appears to be important even in languages like French that maintain voicing during stop closure (van Dommelen, 1983). The relationship between VOT and voicing is, in part, dependent on how contrasts are realized in a particular language. For example, for the same synthetic VOT continuum, Spanish and English speakers have different category boundaries despite the fact that both languages have a single voiced-voiceless contrast (Lisker & Abramson, 1970). In Thai, there are two boundaries, one similar to English and one similar to Spanish, because there is a three way voiced-voiceless-aspirated contrast in the language (Lisker & Abramson, 1970).

Generally, a short or negative VOT is a cue to voicing, a long VOT is a cue to voicelessness, and a very long VOT is a cue to aspiration (in languages with an aspiration contrast). For English, and presumably other languages, the relative amplitude and the presence or absence of aspiration noise is a contributing cue to voicing for word initial stops (Repp, 1979). An additional cue to voicing in syllable onset stops is the relative amplitude of the release burst: a low amplitude burst cues voiced stops while a high amplitude burst cues voiceless stops (Repp, 1977).

The duration and spectral properties of the preceding vowel also provide cues to voicing in post-vocalic stops and fricatives (Soli, 1982). When the vowel is short, with a shorter steady state relative to its offset transitions, voicelessness is perceived. The duration of the consonant stricture is also a cue to both fricative and stop voicing: longer duration cues voicelessness (Massaro & Cohen, 1983). Figure 4 schematically illustrates some of the most frequently cited cues to consonant voicing for two types of

BEST COPY AVAILABLE

consonants: a voiceless alveolar stop /t/, and a voiced alveolar fricative /z/. The horizontal bars represent the first three formants of the vowel /a/, the deformation of the bars represents formant transitions, the hatched areas represents fricative and stop release noises, and the dark bar at the base of the /z/ represents the voicing bar. Table 2 summarizes the consonant manner cues discussed above. Again, the table should not be seen as exhaustive as there are many secondary and contextual cues that contribute to a consonant percept that are not listed here.

Insert Figure 4 about here

Table 3

Summary of Commonly Cited Voicing Cues

Cue	Applies to	Distribution
periodicity	stops, affricates, fricatives	internal
VOT	stops	CV transition
consonant duration	stops, fricatives	internal
release amplitude	stops	C-release
preceding V duration	obstruents	preceding vowel

Visual Information: Multi-modal Speech Perception

Much of the research on speech perception focuses on the acoustic channel alone. In part, the concentration on auditory perception is related to the fact that the acoustic signal is richer in information about spoken language than the visual signal. However, the visual signal may have a large impact on the perception of the auditory signal under degraded conditions. When a hearer can see a talker's face, the gain in speech intelligibility in a noisy environment is equivalent to a 15 dB gain in the acoustic signal alone (Sumby & Pollack, 1954). This is a dramatic difference, superior to that of even the most sophisticated hearing aids. The relative importance of the visual signal increases as the auditory channel is degraded through noise, distortion, filtering, hearing loss, and potentially through unfamiliarity with a particular talker, stimulus set, or listening condition.

When information in the visual channel is in disagreement with the information in the auditory channel, the visual channel may change or even override the percept of the auditory channel alone. McGurk and MacDonald (1976) produced stunning evidence, now known as the "McGurk effect", of the strength of the visual signal in the perception of speech in an experiment that has since been replicated under a variety of conditions. They prepared a video tape of a talker producing two syllable utterances with the same vowel but varying in the onset consonants such as "baba", "mama", or "tata". The audio and video channels were separated and the audio tracks of one utterance were dubbed onto video tracks of different utterances. With their eyes open, subjects' perceptions were strongly influenced by the video channel. For example when presented with a video of a talker saying "tata" together with the audio of the utterance "mama" the subjects perceived "nana". But with their eyes closed, subjects perceived "mama". This effect of cross-modal integration is strong and immediate, there is no hesitation or contemplation on the part of the

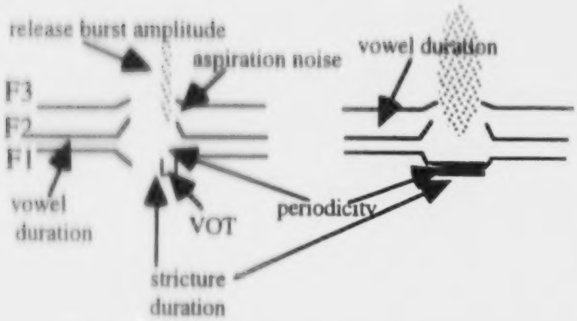

release burst amplitude
aspiration noise
vowel duration
F3
F2
F1
vowel
duration
VOT
periodicity
stricture
duration

Figure 4. Schematic illustration of the voicing cues in two VCV sequences where V is the vowel /a/ and C is a voiceless alveolar stop /t/ and a voiced alveolar fricative /z/.

subjects who are completely unaware of the conflict between the two channels. The McGurk effect is considered by many theorists as evidence that the auditory and visual integration occurs at a 'low' level because of its automatic nature. It also reflects limitations of the information which can be obtained through the visual channel. Many aspects of the speech production process are hidden from view. These include voicing, nasalization, and many vowel and consonant contrasts.

Non-Linearity of the Speech Signal

As a result of the way in which speech is produced, much of the information in the signal is distributed, overlapping, and contextually varying. In producing speech, the articulatory organs of the human vocal tract move continuously with sets of complex gestures that are partially or wholly coextensive and covarying (see the chapter on speech production, this volume). The resulting acoustic and visual signals are continuous and the information that can be identified with a particular linguistic unit shows a high degree of overlap and covariance with information about adjacent units (Delattre, Liberman, & Cooper, 1955; Liberman, Delattre, Cooper, & Gerstman, 1954). This is not to say that segmentation of the signal is impossible; acoustic analysis reveals portions of the signal that can act as reliable acoustic markers for points at which the influence of one segment ends or begins (Fant, 1962). However, the number of segments determined in this way and their acoustic characteristics are themselves highly dependent on the context (Fant, 1962, 1986).

As noted earlier, because of the distributed and overlapping nature of phonetic/linguistic information, the speech signal fails to meet the *linearity* condition (Chomsky & Miller, 1963). This poses great problems for phoneme based speech recognizers and, if discrete units do play a part in perception, they should also be problematic for the human listener. Yet, the listener appears to segment the signal into discrete and linear units such as words, syllables, and phonemes with little effort. In the act of writing, much of the world's population can translate a heard or internally generated signal into word-like units, syllable-like units, or phoneme-like units. Although this is often cited as an argument for a segmentation process in speech perception, the relation between the discrete representation of speech seen in the world's writing systems and the continuous signal is complex and may play little role in the perceptual process (Pierrehumbert & Pierrehumbert 1993). Segmentation may be imposed on an utterance after the perceptual process has been completed. It is not clear that a signal of discrete units would be preferable; the distributed nature of information in the signal contributes to robustness by providing redundant look-ahead and look-back information. Reducing speech to discrete segments could result in a system that, unlike human speech perception, cannot recover gracefully from errorful labeling (Fowler & Smith, 1986; Klatt, 1979, 1990; Pisoni, 1997).

Informational Burden of Consonants and Vowels

From an abstract phonemic point of view, consonant phonemes bear a much greater informational burden than vowel phonemes. That is, there are far more consonant phonemes in English than there are vowel phonemes, and English syllables permit more consonant phonemes per syllable than vowel phonemes. Thus, many more lexical contrasts depend on differences in consonant phonemes than in vowel phonemes. However, the complex overlapping and redundant nature of the speech signal means that the simple information theoretic analysis fails in its predictions about the relative importance of consonant and vowel portions of the signal in speech perception.

The importance of vowels is due to the gross differences in the ways consonants and vowels are produced by the vocal tract. Consonants are produced with a complete or partial occlusion of the vocal tract, causing a rapid attenuation of the signal, particularly in the higher frequencies. In the case of oral stops, all but the lowest frequencies (which can emanate through the fleshy walls of the vocal tract) are absent from the signal. In contrast, vowels are produced with a relatively open vocal tract, therefore there is little overall attenuation and formant transitions are saliently present in the signal (Fant, 1960). This dichotomy means that the vowels are more robust in noise and that vowel portions of the signal carry more information about the identity of the consonant phonemes than the consonant portions of the signal carry about the vowel phonemes.

Although they are partially the result of articulator movement associated with consonants, the formant transitions are considered part of the vowel because of their acoustic characteristics. Generally speaking, the transitions have a relatively high intensity and long duration compared to other types of consonantal cues in the signal. The intensity, duration, and periodic structure of the transitions make them more resistant to many types of environmental masking than release bursts, nasal pole-zero patterns, or frication noise. Formant transitions bear a dual burden of simultaneously carrying information about both consonant and vowel phonemes. In addition, information about whether or not a consonant phoneme is a nasal, a lateral, or a rhotic is carried in the vowel more effectively than during the consonant portion of the signal. Figure 5 is a speech spectrogram of the word "formant" illustrating the informational burden of the vowel. What little information that consonants carry about the flanking vowel phonemes is found in portions of the signal that are low intensity, aperiodic, or transient. Therefore, it is more easily masked by environmental noise.

Insert Figure 5 about here

It is well known that spoken utterances are made up of more than the segmental distinctions represented by consonant and vowel phonemes. Languages like English rely on lexical stress to distinguish words. The majority of the world's languages have some form of *tone* contrast, whether fixed on a single syllable as in the Chinese languages, or mobile across several syllables as in Kikuyu (Clements, 1984). *Pitch-accent*, like that seen in Japanese, is another form of tone based lexical distinction. Tone and pitch-accent are characterized by changes in voice pitch (fundamental frequency) and in some cases changes in voice quality such as creakiness or breathiness (as in Vietnamese, Nguyen, 1987). These types of changes in the source function are carried most saliently during the vowel portions of the signal. Although stress affects both consonant and vowels, it is marked most clearly by changes in vowel length, vowel formants, and in fundamental frequency excursions. Prosodic information is carried by both consonants and vowels; however, much of it takes the form of pitch, vowel length, and quality changes. Thus, despite the relatively large information burden that consonant phonemes bear, the portions in the physical signal that are identified with vowels carry much more of the acoustic information in a more robust fashion than the portions of the signal associated with the consonants.

Invariance and Variability

In addition to violating the linearity condition, the speech signal is characterized by a high degree of variability, violating the invariance condition. There are many sources of variability that may be interrelated or independent. Variability can be broken into two broad categories: 1) production related and 2) production independent. It is worth noting that while production related variability is complex, it is 'lawful' and is a potentially rich source of information both about the intended meaning of an utterance and about the talker. Production independent variability derives from such factors as environmental noise or reverberation and may provide the listener with information about the environmental conditions surrounding the conversation; it can be seen as random in its relation to the linguistic meaning of the utterance and to the talker. Understanding how the perceptual process deals with these different types of variability is one of the most important issues in speech perception research. In traditional symbol-

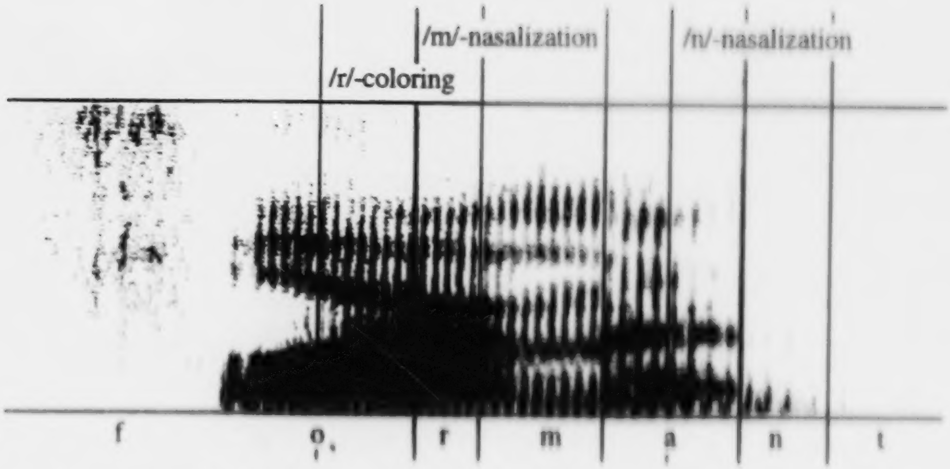


Figure 5. A spectrogram of the word "formant" illustrating the information that adjacent vowels carry about rhotics (/r/-coloring) and nasals (nasalization).

processing approaches that treat variation as noise, listeners are thought to compensate for differences through a processes of perceptual *normalization* in which linguistic units are perceived relative to the context, e.g., the prevailing rate of speech (e.g., Miller, 1981, 1987; Summerfield, 1981) or the dimensions of the talker's vocal tract (e.g., Joos, 1948; Ladefoged & Broadbent, 1957; Summerfield & Haggard, 1973). Alternative non-analytic approaches to speech perception that are based on episodic memory (Goldinger 1997, Johnson 1997, Pisoni 1997) propose that speech is encoded in a way that preserves the fine details of speech production related variability. While these approaches may use some types of variability in the speech perception process, little has been said about the production-independent variability. The following section, while not exhaustive, is a sampling of some well known sources of variability and their impact on speech perception (for a more detailed discussion see Klatt 1975, 1976, 1979). Production related variability in speech applies both across talkers as a result of physiological, dialectal, and socioeconomic factors, as well as within a talker from one utterance to the next as a result of factors such as coarticulation, rate, prosody, emotional state, level of background noise, distance between talker and hearer, and semantic properties of the utterance. What follows is a review of some of the most important sources of variability in speech and their effects on perceptual processes.

Coarticulation

The most studied source of within-talker variability, *coarticulation*, is one source of non-linearity in speech. In the production of speech, the gestures in the vocal tract are partially or wholly overlapping in time, resulting in an acoustic signal in which there is considerable contextual variation (Delattre, Liberman, & Cooper, 1955; Liberman, 1957; Liberman, Delattre, Cooper, & Gerstman, 1954). The degree to which any one speech gesture is affected or affects other gestures depends on the movements of the articulators and the degree of its constriction as well as factors such as rate of speech and prosodic position. Although coarticulation is often described as a universal physiological aspect of speech, there is evidence for talker specific variation in the production and timing of speech gestures and in the resulting characteristics of coarticulation (Johnson, Ladefoged & Lindau, 1993; Kuehn & Moll, 1976; Stevens, 1972).

The perceptual problems introduced by coarticulatory variation became apparent early in the search for invariant speech cues. Because of coarticulation, there is a complex relationship between acoustic information and phonetic distinctions. In one context, an acoustic pattern may give rise to one percept, while in another context the same acoustic pattern may give rise to a different percept (Liberman et al., 1954). At the same time, many different acoustic patterns may cue a single percept (Hagiwara, 1995).

Speaking Rate

Changes in speech rate are reflected in changes in the number and duration of pauses, in durational changes of vowels and some consonants, and in deletions and reductions of some of the acoustic properties that are associated with particular linguistic units (J.L. Miller, Grosjean, & Lomato, 1984). For example, changes in VOT and the relative duration of transitions and vowel steady-states occur with changes in speaking rates (J.L. Miller & Baer, 1983; J.L. Miller, Green, & Reeves, 1986; Summerfield, 1975).

There is now a large body of research on the consequences of rate based variability on the perception of phonemes. These findings demonstrate that listeners are sensitive to rate based changes that are internal or external to the target word. The importance of token-internal rate sensitivity was demonstrated by J.L. Miller and Liberman (1979). Listeners were presented with a synthetic /ba/ - /wa/ continuum that varied the duration of the formant transitions and the duration of the vowel. The results showed that the crossover point between /b/ and /w/ was dependent on the ratio of the formant transition duration to the vowel duration: the longer the vowel, the longer the formant transitions had to be to produce the /wa/ percept. The importance of token-external rate sensitivity was demonstrated in an experiment on the identification of voiced and voiceless stops. Summerfield (1981) presented listeners with a precursor phrase that varied in speaking rate followed by a stimulus token. As the rate of the precursor phrase increased, the voiced-voiceless boundary shifted to shorter voice onset time (VOT) values. Sommers, Nygaard, and Pisoni (1992) found that the intelligibility of isolated words presented in noise was affected by the number of speaking rates that were used to generate the test stimulus ensemble: stimuli drawn from three rates (fast, medium, and slow) were identified more poorly than stimuli from only a single speaking rate.

Prosody

Rate based durational variation is compounded by many other factors, including the location of syntactic boundaries, prosody, and the characteristics of adjacent segments (Klatt, 1976; Lehiste, 1970; Pierrehumbert & Beckman, 1988; Beckman & Edwards, 1994). It is well known that lexical stress has a dramatic effect on the articulations that produce the acoustic signal. However, lexical stress is only one level of prosodic hierarchy spanning the utterance. Prosody is defined by Beckman and Edwards (1994, p. 8) as "the organizational framework that measures off chunks of speech into countable constituents of various sizes." Different positions within a prosodic structure lead to differences in articulations which in turn lead to differences in the acoustic signal (Lehiste, 1970; Beckman & Edwards, 1994; Fujimura, 1990; Fougeron & Keating, 1997). For example, vowels that are in the nuclear accented syllable of a sentence ('primary sentential stress') have a longer duration, a higher amplitude, and have a more extreme articulator displacement than vowels in syllables that do not bear nuclear accent (Beckman & Edwards, 1994; de Jong, 1995).

Articulations that are at the edges of prosodic domains also undergo systematic variation which result in changes in the acoustic signal such as lengthened stop closures, greater release burst amplitude, lengthened VOT, and less vowel reduction. These effects have been measured for word-initial versus non-initial positions (e.g., Browman & Goldstein, 1992; Byrd, 1996; Cooper, 1991; Fromkin, 1965; Vassière, 1988, Krakow, 1989) and at phrase and sentence edges (e.g., Fougeron & Keating, 1996). Finally, the magnitude of a local effect of a prosodic boundary on an articulation interacts in complex ways with global trends that apply across the utterance. One such trend, commonly referred to as *declination*, is for articulations to become less extreme and for fundamental frequency to fall as the utterance progresses (e.g., Vassière, 1986; Varya & Fowler, 1992). Another global trend is for domain edge effects to apply with progressively more force as the edges of progressively larger domains are reached (Klatt, 1975; Wightman et al., 1992; Jun, 1993). These factors interact with local domain edge effects in a way that indicates a nested hierarchical prosodic structure (Jun, 1993; Maeda, 1976). However, the number of levels and the relative strength of the effect may be a talker dependent factor (Fougeron & Keating, 1997).

Semantics and Syntax

In addition to prosodic structure, the syntactic and semantic structure have substantial effects on the fundamental frequency, patterns of duration, and relative intensities of vowels (Klatt, 1976; Lehiste, 1967; Lieberman, 1963). For example, when a word is uttered in a highly predictable semantic and syntactic position, it will show a greater degree of vowel reduction (*centralization*), with lower amplitude and a shorter duration than the identical word in a position with low contextual predictability (Fowler & Housman, 1987; Lieberman, 1963). These production differences are correlated with speech intelligibility; if the two words are isolated from their relative contexts, the word from the low-predictability context is more intelligible than the word from the high-predictability context. This type of effect is hypothesized to be the result of the talker adapting to the listener's perceptual needs (Lindblom, 1990; Fowler, 1986; Fowler & Housman, 1987): the more information the listener can derive from the conversational context, the less effort a talker needs to spend maintaining the intelligibility of the utterance. The reduced speech is referred to as 'hypo-articulated' and the non-reduced speech is referred to as 'hyper-articulated' (Lindblom, 1990). Similar patterns of variability can seen in many other production related phenomena such as the Lombard reflex (described below). This variability interacts with other factors like speaking rate and prosody in a complex fashion, making subsequent normalization extremely difficult. Yet, listeners are able to extract and use syntactic, semantic, and prosodic information from the lawful variability in the signal (Klatt, 1976; McClelland & Elman, 1986).

Environmental Conditions

There are many factors that can cause changes in a talker's source characteristics and in the patterns of duration and intensity in the speech signal that are not directly related to the talker or to the linguistic content of the message. These include the relative distance between the talker and the hearer, the type and level of background noise, and transmission line characteristics. For example, in a communicative situation in which there is noise in the environment or transmission line, there is a marked rise in amplitude of the produced signal that is accompanied by changes in the source characteristics and changes the dynamics of articulatory movements which together are known as the "Lombard reflex" (Cummings, 1995; Gay, 1977; Lombard, 1911; Lane & Tranel, 1971; Schulman, 1989). The Lombard reflex is thought to result from the need of the talker to maintain a sufficiently high signal-to-noise ratio to maintain intelligibility.

As the environmental conditions and the distance between talker and hearer are never identical across instances of any linguistic unit, it is guaranteed that no two utterances of the same word in the same syntactic, semantic, and prosodic context will be identical. Furthermore, in natural settings the level of environmental noise tends to vary continuously so that even within a sentence or word, the signal may exhibit changes. Similarly, if the talker and listener have the ability to communicate using both the visual and auditory channels, the resulting speech signal exhibits selective reductions such as those seen for high semantic context or good signal to noise ratios; but when there is no visual channel available, the resultant speech is marked by hyper-articulation that is similar to that seen for the Lombard reflex or for low semantic predictability contexts (Anderson, Sotillo, & Doherty-Sneddon, 1997). Like other types of hypo- and hyper-articulation, the variation based on access to visual information is highly correlated with speech intelligibility.

Physiological Factors

Among the most commonly cited sources of between talker variation are differences in the acoustic signal based on a talker's anatomy and physiology. The overall length of the vocal tract and the relative size of the mouth cavity versus the pharyngeal cavity determines the relative spacing of the formants in vowels (Chiba & Kajiyama, 1941; Fant, 1960, 1973). These differences underlie some of the male-female and adult-child differences in vowels and resonant consonants (Fant, 1973; Bladon, Henton, & Pickering, 1984; Hagiwara, 1995). Moreover, vocal tract length may also contribute to observed differences in obstruent voicing (Flege & Massey, 1980) and fricative spectra (Schwartz, 1968; Ingemann, 1968). Physiological differences between male, female and child laryngeal structure also contribute to observed differences in the source characteristics such as fundamental frequency, spectral tilt and noisiness (Henton & Bladon, 1985; Klatt & Klatt, 1990). Other types of physiologically based differences among talkers that are expected to have an effect on the acoustic signal include dentition, size and doming of the hard palate, and neurological factors such as paralysis or motor impairments.

The importance of talker specific variability in the perception of linguistic contrasts was first reported by Ladefoged and Broadbent (1957). Listeners were presented with a precursor phrase in which the relative spacing of the formants was manipulated to simulate established differences in vocal tract length. The stimulus was one of a group of target words in which the formant spacing remained fixed. The listeners' percepts were shifted by the precursor sentence. In a follow-up experiment, a different group of listeners was presented with the same set of stimuli under a variety of conditions and instructions (Ladefoged, 1967). Even when listeners were told to ignore the precursor sentence or when the sentence and the target word were presented from different loudspeakers, the vowel in the target word was reliably shifted by the formant manipulation of the precursor sentence. This effect was only successfully countered by placing the target word before the sentence or by having the listeners count aloud for 10 seconds between the precursor and hearing the target word.

Dialectal and Ideolectal Differences

In addition to physiologically based differences, there are a number of socioeconomic and regional variables that affect the production of speech. Perception of speech from a different dialect can be a challenging task. Peterson and Barney (1952) found differences in dialect to be one of the most important sources of confusions in the perception of vowel contrasts. Research on improving communications reliability found that training talkers to avoid dialectal pronunciations in favor of 'Standard English' was much easier than training listeners to adapt to a variety of dialects (Black & Mason, 1946). Differences between individual speakers' styles, or idiolect, also require a certain amount of adaptation on the part of the listener. Dialectal and ideolectal variability have received relatively little attention in the speech perception literature and are generally treated as additional sources of noise which are discarded in the process of normalization.

Robustness of Speech Perception

In everyday conversational settings there are many different sources of masking noise and distortions of the speech signal, yet only under the most extreme conditions is perceptual accuracy affected. Much of the robustness of speech comes from the redundant information which is available to the listener. Since the main goal of speech is the communication of ideas from the talker to the hearer (normally under less than optimal conditions), it is not surprising that spoken language is a highly redundant system of information transmission. While redundancy of information in a transmission system implies inefficient encoding, it facilitates error correction and recovery of the intended signal in a noisy environment and insures that the listener recovers the talker's intended message.

The redundancy of speech resides, in part, in the highly structured and constrained nature of human language. Syntactic and semantic context play a large role in modulating the intelligibility of speech. Words in a sentence are more predictable than words spoken in isolation (Fletcher, 1929; French & Steinberg, 1947; Miller, 1962; Miller, Heise, & Lichten, 1951; Pollack & Pickett, 1964). The sentence structure (syntax) of a particular language restricts the set of possible words that can appear at any particular point in the sentence to members of appropriate grammatical categories. The semantic relationships between words also aids in perception by further narrowing the set of words that are likely to appear in a sentence. It has been shown experimentally that limiting the set of possible words aids in identification. For example, Miller et al. (1951) found that limiting the vocabulary to digits alone results in an increase in speech intelligibility. More generally, lexical factors like a word's frequency of usage and the number of acoustically similar words have been shown to have a dramatic impact on a word's intelligibility (Anderson, 1962; Luce, 1986; Treisman, 1978).

Phonological structure also constrains the speech signal and facilitates the listener's perception of the intended message. Prosodic structure and intonation patterns provide auditory cues to syntactic structure, which reduces the number of possible parses of an utterance. The syllable structure and stress patterns of a language limit the number of possible speech sounds at any particular point in an utterance, which aids in identifying words (see for example Cutler, 1997; Norris & Cutler, 1995).

Much of the 'top down' information in language is contextual in nature and resides in the structural constraints on a given language and not in the speech signal itself (Jelinek 1998). However, because prosodic, syntactic and semantic factors create systematic variability in production (discussed in the variability section), the signal contains a significant amount of information about the linguistic structures larger than the segment, syllable, and word. While the role of suprasegmental information (above the level of the phoneme) has traditionally received less attention in the perception literature, there have been a few studies that reveal the richness of suprasegmental information in the speech signal.

In spectrogram reading experiments Cole, Rudnicky, Reddy, & Zue (1978) demonstrated that the acoustic signal is rich in information about the segmental, lexical, and prosodic content of an utterance. An expert spectrogram reader who was given the task of transcribing an utterance of unknown content using speech spectrograms alone achieved an 80-90 percent accuracy rate. This finding demonstrates that not only are features that cue segmental contrasts present in the signal, but prosodic and word boundary information is also available as well. However, it is not clear from these spectrogram reading experiments whether the features that the transcriber used are those that listeners use. A study that tests the ability of listeners to draw on the richness of the signal was conducted by Liberman and Nakatani (cited in Klatt, 1979). Listeners who were given the task of transcribing pseudo-words embedded in normal English sentences achieved better than 90 percent accuracy.

There are numerous other studies that demonstrate the importance of prosodic melody (e.g., Collier & t'Hart, 1975; Klatt & Cooper, 1975) in sentence and word parsing. For example, Lindblom and Svensson (1973), using stimuli in which the segmental information in the signal was removed, found that listeners could reliably parse sentences based on the prosodic melody alone. Prosody has been found to play a role in perceptual coherence (Darwin, 1975; Studdert-Kennedy, 1980) and to play a central role in predicting words of primary semantic importance (e.g., Cutler, 1976).

A second source of the redundancy in speech comes from the finding that the physical signal is generated by the vocal tract. As we have already noted, speech sounds are overlapped, or coarticulated, when they are produced, providing redundant encoding of the signal. The ability to coarticulate, and thereby provide redundant information about the stream of speech sounds serves to both increase transmission rate (Liberman 1996) and provide robustness to the signal (Wright 1996). Redundancy in the acoustic signal has been tested experimentally by distorting, masking, or removing aspects of the signal and exploring the effect these manipulations have on intelligibility. For example, connected speech remains highly intelligible when the speech power is attenuated below 1800 Hz or when it is attenuated above 1800 Hz (French & Steinberg, 1947). This finding indicates that speech information is distributed redundantly across lower and higher frequencies. However, not all speech sounds are affected equally by frequency attenuation. Higher frequency attenuation causes greater degradation for stop and fricative consonants, while lower frequency attenuation results in greater degradation of vowels, liquids (/r/ and /l/ in English), and nasal consonants (Fletcher, 1929). For example, the place of articulation distinctions among fricatives are carried in large part by the fricative noise, which tends to be concentrated in higher frequencies. Attenuating these particular frequencies results in an increase in fricative confusions and a decrease in intelligibility.

Speech can be distorted in a natural environment by reverberation. Experiments on the perception of nonsense syllables found that intelligibility was relatively unaffected by reverberation with a delay of less than 1.5 seconds. Reverberation with a greater delay caused a marked drop off in intelligibility (e.g., Knudsen, 1929; Steinberg, 1929). Under extremely reverberatory conditions, individual speech sounds blend together as echoes overlap in a way that causes frequency and phase distortions. Again, not all speech sounds are equally affected by reverberation. Long vowels and fricatives, which have an approximately steady state component, are much less susceptible to degradation than short vowels and nasal and stop consonants, which are distinguished from each other by relatively short and dynamic portions of the signal (e.g., Steinberg, 1929).

Overall, the intelligibility of individual speech sounds in running speech is in part a function of their intensity (Fletcher, 1929). In general, vowels are more intelligible than consonants. More specifically, those consonants with the lowest intensity have the poorest intelligibility. Of these, the lease reliably identified in English are the non-sibilant fricatives, such as those found in "fat", "vat", "thin", and "this". These fricatives achieve 80 percent correct identification only when words are presented at relatively high signal to noise ratios (Fletcher, 1929). These fricatives' noises are also spectrally similar, adding to their confusability with each other. English is one of the few languages that contrasts non-sibilant fricatives (Maddieson, 1984), presumably due to their confusability and low intelligibility. By contrast, the sibilant fricatives (for example those found in the words "sap", "zap", "Confucian", and "confusion") have a much greater intensity and are more reliably identified in utterances presented at low signal to noise ratios. The next most intelligible sounds are the stop consonants, including /p/, /t/, and /k/ in English, followed by the vocalic consonants such as the nasals and liquids. Vowels are the most identifiable. The low vowels, such as those found in the words "cot" and "caught", are more easily identified than the high vowels, such as those found in "peat" and "pit".

Models and Theories

Theories of human speech perception can be divided into two broad categories, those that attempt to model segmentation of the spoken signal into linguistic units (which we refer to as models of speech perception) and those which take as input a phonetic transcription and model the access of the mental lexicon (which we refer to as models of spoken word recognition). Almost all models of speech perception try to identify phonemes in the signal. A few models go straight to the word level, and thus encompass the process of word recognition as well. These models are discussed in the section on word recognition below.

Models of Human Speech Perception

Most current models of speech perception have as their goal the segmentation of the spoken signal into discrete phonemes. All of the models discussed in this section have this property. In addition, these models assume that at some level of perceptual processing there are invariant features which can be extracted from the acoustic signal, though which aspects are taken to be invariant depends on the model.

Invariance Approaches

The most extensively pursued approach to solving the variability problem is the search for invariant cues in the speech signal. This line of research which dates back to the beginning of modern speech research in the late 1940s has revealed a great deal of coarticulatory variability. It has resulted in a series of careful and systematic searches for invariance in the acoustic signal that has revealed a wealth of empirical data. Although researchers investigating acoustic-phonetic invariance differ in their approaches, they have in common the fundamental assumption that the variability problem can be resolved by studying more sophisticated cues than were originally considered (e.g., Blumstein & Stevens, 1980; Fant, 1967; Mack & Blumstein, 1983; Kewley-Port, 1983). Early experiments on speech cues in speech perception used copy-synthesized stimuli in which much of the redundant information in the signal had been stripped away. In addition, acoustic analysis of speech using spectrograms focused only on gross characteristics of the signal.

One approach, termed *static* (Nygaard & Pisoni, 1995), is based on the acoustic analysis of simple CV syllables. This approach focused on complex integrated acoustic attributes of consonants that are hypothesized to be invariant in different vowel contexts (e.g., Blumstein & Stevens, 1979). Based on Fant's (1960) acoustic theory of speech production, Stevens and Blumstein (1978, 1981; also Blumstein & Stevens, 1979) hypothesized invariant relationships between the articulatory gestures and acoustic features that are associated with a particular segment. They proposed that the gross spectral shape at the onset of the consonant release burst is an invariant cue for place of articulation. In labial stops (/p/ and /b/), the spectral energy is weak and diffuse with a concentration of energy in the lower frequencies. For the alveolar stops (/t/ and /d/) the spectral energy is strong but diffuse with a concentration of energy in the higher frequencies (around 1800 Hz). Velar stops (/k/ and /g/) are characterized by strong spectral energy that is compact and concentrated in the mid-frequencies (around the 1000 Hz).

A different approach, termed *dynamic* (Nygaard & Pisoni, 1995), has been proposed by Kewley-Port (1983). She employed auditory transformations of the signal, looking for invariant dynamic patterns in running spectra of those transformations. The dynamic approach is promising because it can capture an essential element of the speech signal: its continuous nature. More recent static approaches adopted an element of dynamic invariance into their approaches (Mack & Blumstein, 1983; Lahiri, Gewirth, & Blumstein, 1984).

As is noted by Nygaard & Pisoni (1995), any assumption of invariance necessarily constrains the types of processes that underlie speech perception. Speech perception will proceed in a bottom up fashion with the extraction of invariant features or cues being the first step in the process. Invariance explicitly assumes abstract canonical units and an elimination of all forms of variability and noise from the stored representation (see Pisoni 1997) This includes many sources of variation that are potentially useful to the listener in understanding an utterance. For example, indexical and prosodic information is discarded in the reduction of the information in the signal to a sequence of idealized symbolic linguistic invariants.

While these approaches to speech sound perception have provided some promising candidates for extraction of invariant features from the signal (Sussman, 1989) and have produced invaluable empirical data on the acoustic structure of the speech signal and its auditory transforms, they have done so for only a very limited set of consonants in a very limited set of contexts. For example, the three places of articulation treated by Stevens and Blumstein represent slightly less than one quarter of the known consonant places of articulation in the world's languages (Ladefoged & Maddieson, 1996). Even for the same places of articulation, the features found in English may not invariantly classify segments of other languages (Lahiri, Gewirth, & Blumstein, 1984). Furthermore, much of the contextual variability that is non-contrastive in English, and therefore removed in the invariance approach, forms the basis for a linguistic contrast in at least one other language. Therefore, the type of processing that produces invariant percepts must be language specific.

Motor Theory

One of the ways in which the perception of speech differs from many other types of perception is that the perceiver has intimate experience in the production of the speech signal. Every listener is also a talker. The motor theory (Liberman et al., 1967) and the revised motor theory (Liberman & Mattingly, 1986, 1989) take advantage of this link by proposing that perception and production are related by a common set of neural representations. Rather than looking for invariance in the acoustic signal, the perceiver is hypothesized to recover the underlying intended phonetic gestures from an impoverished and highly encoded speech signal. The intended gestures of the talker are therefore assumed to be perceived directly via an innate phonetic module conforming to the specifications of modularity proposed by Fodor

(1983). The phonetic module is proposed to have evolved for the special purpose of extracting intended gestures preemptively (Liberman, 1982; Mattingly & Liberman, 1989; Whalen & Liberman, 1987). That is, the phonetic module gets first pass at the incoming acoustic signal and extracts the relevant phonetic gestures passing the residue on for general auditory processing (see also Gaver 1993).

A variety of experiments showing that speech is processed differently from non-speech provide evidence for a neural specialization for speech perception. Some of these findings have subsequently been shown to apply equally as well to non-speech stimuli (see for example: Pisoni 1977; Jusczyk, 1980; Eimas & Miller, 1980; Repp, 1983a, b; for a review see Goldinger, Pisoni, & Luce, 1996). While some evidence for the specialness of speech still stands, it is uncertain whether appropriate non-speech controls to compare to speech have been considered. A number of ways of creating complex signals which are more or less acoustically equivalent to speech have been considered; however, these experiments do not explore whether there are controls which are communicatively or informationally equivalent to speech.

A good example of the importance of testing the evidence with informationally equivalent stimuli can be found in a phenomenon known as duplex perception (Rand, 1974) which has been cited frequently as strong evidence for a speech specific module (Liberman, 1982; Repp, 1982; Studdert-Kennedy, 1982; Liberman & Mattingly, 1989). To elicit duplex perception, two stimuli are presented dichotically to a listener wearing headphones. An isolated third formant transition, which sounds like a chirp, is presented in one ear while the "base syllable", which is ambiguous because it has had the third formant transition removed, is presented in the other ear. The isolated formant transition fuses with the base syllable, which is then heard as an unambiguous syllable in the base ear. Additionally, the chirp is perceived separately in the other ear. Duplex perception was found to occur with speech stimuli but not with acoustically equivalent stimuli. However, the informational equivalence of the stimuli was brought into question by Fowler and Rosenblum (1990) who found that a natural sound, the sound of a door slamming, patterned more like speech in a duplex perception task, and differently from laboratory generated non-speech controls (which are complex artificial sound patterns). A door slam is ecologically relevant, as it gives the hearer information about an action which has occurred in the world (see also Pastore, Schmuckler, Rosenblum, & Szczesiul, 1983; Nusbaum, Schwab, & Sawusch, 1983; Gaver, 1993). Speech has tremendous social significance and is probably the most highly practiced complex perceptual task performed by humans. These factors have not been adequately considered when explaining differences between speech and non-speech perception.

A claim of the original formulation of the motor theory (Liberman et al., 1967) was that the percepts of speech are not the acoustic signals which impinge directly upon the ear, but rather the articulations made by the speaker. One of the striking findings from early experiments was that the discontinuities in the acoustic-to-phonemic mapping for stop onset consonants (Liberman, Delattre, & Cooper, 1952). These discontinuities were taken as crucial evidence against an acoustic basis for phoneme categories. However, researchers have found that for some phonemic categories the acoustic mapping is simple while the articulatory mapping is complex. For example, American English /r/ can be produced with one or more of three distinct gestures, and there is intraspeaker variation in which different gestures are used (Delattre & Freeman 1968; Hagiwara 1995; see also Johnson, Ladefoged, & Lindau 1993).

The search for first-order acoustic invariance in speech has been largely unsuccessful, and it is now well known that the articulatory gestures and even their motor commands are not invariant either (e.g., MacNeilage, 1970). In the revised motor theory, the articulatory percepts are assumed to be the speaker's "intended" gestures, before contextual adjustments and other sources of speaker independent variability in production (Liberman & Mattingly, 1986, 1989). Thus, in terms of the nature of neural representations, the motor theory's proposed linguistic representations are extremely abstract, canonical symbolic entities that can be treated as formally equivalent to abstract phonetic segments. Since neither acoustic nor articulatory categories provide simple dimensions upon which to base perceptual categories in speech, as the coherence as categories of these abstractions can be based on either articulatory or acoustic properties, or both.

There are several appealing aspects of the motor theory of speech perception. It places the study of speech perception in an ecological context by linking production and perception aspects of spoken language. It also accounts for a wide variety of empirical findings in a principled and consistent manner. For example, the McGurk effect can be nicely accommodated by a model that is based on perception of gestures, although direct perception (Fowler, 1986; Fowler & Rosenblum, 1991) and FLMP (Oden & Massaro, 1978; Massaro & Cohen, 1993; Massaro, 1989) also incorporate visual information, although in very different ways. Despite the appeal of the motor theory, there remain several serious shortcomings. The proposed perceptual mechanisms remain highly abstract, making effective empirical tests of the model difficult to design. A more explicit model of how listeners extract the intended gestures of other talkers would go far to remedy this problem. In addition, the abstract nature of the intended gestures involves a great deal of reduction of information and therefore suffers from the same shortcomings that traditional phonemic reduction does: it throws away much of the inter- and intra-talker variability which is a rich source of information to the listener.

Direct-Realist Approach

The direct-realist approach to speech perception (Fowler, 1990; Fowler & Rosenblum, 1991) draws on Gibson's (1966) ecological approach to visual perception. Its basic assumption is that speech perception, like all other types of perception, acts directly on events in the perceiver's environment rather than on the sensory stimuli and takes place without the mediation of cognitive processes. An event may be described in many ways but those that are ecologically relevant to the perceiver are termed 'distal events'. The sets of possibilities for interaction with them are referred to as 'affordances'. The distal event imparts structure to an informational medium, the acoustic signal and reflected light in the case of visible speech, which in turn provides information about the event to the perceiver by imparting some of its structure to the sense organs through stimulation. The perceiver actively seeks out information about events in the environment, selectively attending to aspects of the environmental structure (Fowler, 1990).

In speech, the phonetically determined coordinated set of movements of the vocal tract that produce the speech signal are the events that the perceiver is attending to. In this way, the direct realist approach is like the motor theory. However, rather than assuming a speech specific module retrieving intended gestures from an impoverished acoustic signal, the direct realist approach assumes an information rich signal in which the phonetic events are fully and uniquely specified. Because the perception is direct, the direct realist approach views variability and nonlinearity in a different light than most other approaches to speech perception which are abstractionist in nature.

The vocal tract cannot produce a string of static and non-overlapping shapes, so the gestures of speech cannot take place in isolation of each other. Direct perception of gestures gives the listener detailed information about both the gestural and environmental context. This implies that the perceiver is highly experienced with the signal and so long as that variation is meaningful, it provides information about the event. Rather than removing noise through a process of normalization, variation provides the perceiver with detailed information about the event which includes the talker's size, gender, dialect region, emotional state, as well as prosodic and syntactic information. Therefore, according to this view, stimulus variation ceases to be a problem of perception and becomes a problem of perceptual organization. While direct perception focuses on the perceived events as gestural constellations roughly equivalent to the phoneme, it is also compatible with the theory to assume the perceived events are words. Thus, we might also consider direct perception as a model of spoken word recognition. Direct perception shares with the exemplar models (discussed in the word recognition section) the assumption that the variability in the signal is rich in information which is critical to perception.

Direct perception is appealing because of its ability to incorporate and use stimulus variability in the signal, and because it makes the link between production and perception transparent. However, there are several important theoretical issues that remain urresolved. One potential problem for a model that permits no mediation of cognitive processes are top down influences on speech perception. As was noted previously, these effects are extremely robust and include phoneme restoration (Samuel, 1981; Warren, 1970), correction of errors in shadowing (Marslen-Wilson & Welsh, 1978), mishearings (Browman, 1978; Garnes & Bond, 1980), lexical bias (Ganong, 1980), syntactic and semantic bias (Salasoo & Pisoni, 1985), and lexical frequency and density bias. Fowler (1986) acknowledges this problem and suggests that there may be special mechanisms for highly learned or automatic behavior and for perceivers' hypothesizing information that is not detected in the signal. She suggests that while perception itself must be direct, behavior may often not be directed by perceived affordances. In this way, the direct-realist perspective departs dramatically from other versions of event perception (Gibson, 1966) which have nothing to say about cognitive mediation.

Finally, Remez (1986) notes that it is not clear that the perceptual objects in linguistic communication are the gestures which create the acoustic signal. While visual perception of most objects is unambiguous, speech gestures are very different in terms of their perceptual availability (Diel, 1986; Porter, 1986; Remez, 1986). Fowler proposes that the perception of the articulatory gestural complex is the object of perception, but the articulations themselves are a medium that is shaped by the intended linguistic message. As she notes herself, while visually identified objects are perceived as such, listeners intuitions are that they perceive spoken language not as a series of sound producing actions (i.e., gestures) but as a sequence of words and ideas. This difference is not necessarily a problem for the model itself, but rather a problem for the way this approach has thus far been employed.

FLMP

A radically different approach to speech perception is represented by informational models which are built around general cognitive and perceptual processes. Most of these models have been developed to explain phonemic perception, and they typically involve multiple processing stages. One example of this approach is the fuzzy logic model of perception, or FLMP (Oden & Massaro, 1978; Massaro & Cohen, 1993; Massaro, 1989). FLMP was developed to address the problem of integrating information from multiple sources, such as visual and auditory input, in making segmental decisions. The criterion for perception of a particular set of features as a particular perceptual unit such as the phoneme is goodness of the percept's match to a subjectively derived prototype description in memory, arrived at through experience with the language of the listener. In acoustic processing, the speech signal undergoes an acoustic analysis by the peripheral auditory system. Evidence for phonemic features in the signal are evaluated by feature detectors using continuous truth values between 1 and 0 (Zadeh, 1965). Then feature values are integrated and matched against the possible candidate prototypes. Because fuzzy algorithms are used, an absolute match is not needed for the process to achieve a phonemic percept.

There are several aspects of FLMP which make it an appealing model. The first is that it provides an explicit mechanism for incorporating multiple sources of information from different modalities. This is particularly important considering the role that visual input can play in the speech perception process. Second, it provides a good fit to data from a wide variety of perceptual experiments (see Massaro, 1987 for a review). Third, it is one of the only models of speech perception that is mathematically explicit, because it is based on a precise mathematical framework (Townsend, 1989). However, there are several serious shortcomings to the model. The most severe, noted by Klatt (1989), Repp (1987) and others is that it is unclear that the fuzzy values are flexible enough to account for the variation that is observed in the speech signal. Because the model works with features to be matched to stored prototypes in memory, there is still a reliance on exclusivity of invariant features and the dependence of features on a degree of normalization across the many sources of variability observed in conversational speech. Moreover, the model has no connection to the perception-production link. Finally, FLMP employs a large number of free parameters that are deduced from the data of specific experimental paradigms but which do not transfer well across paradigms.

Models of Spoken Word Recognition

Models of spoken word recognition can be broken down into two types: those that act on a phonemic or broad phonetic representations, and those that work directly on the acoustic input. Models based on a phonemic level are inspired, or transparently derived, from models of alphabetic reading. Because these models use a unitized input, they explicitly or implicitly assume access to a phonemic or featural representation. These models require either an additional preprocessor which recognizes phonemes, segments, or features, or they assume direct perception of these units from information in the acoustic signal. Models which work on segmental or featural input are by far the most numerous and best known, and only a few that are representative of the diversity of proposals will be discussed here. These are TRACE, NAM, and SHORTLIST. Models that act on the speech signal, or an auditory transformation thereof, necessarily incorporate the speech perception process into the word recognition process. Of the few models of this type, two examples will be discussed: LAFS and Exemplar-covering models.

TRACE

The TRACE model (Elman, 1989; Elman & McClelland, 1986; McClelland & Elman, 1986) is an example of an interactive activation/competition connectionist model. The most widely discussed version of TRACE takes allophonic level features as their input. An early form of the model (Elman and McClelland, 1986a) takes the speech signal as its input and relies on feature detectors to extract relevant information; however, this version was quite limited, being built around only nine CV syllables produced by a single talker.

TRACE is constructed of three levels representing features, phonemes and words. The featural level passes activation to the phonemic level which in turn passes activation to the word level. Within each level, the functional units are highly interconnected nodes each with a current activation level, a resting level, and an activation threshold. There are bi-directional connections between units of different levels and between nodes within a level. Connections are excitatory between units at different levels that share common properties (e.g., between voice, place and manner features and a particular consonantal phoneme). Connections between units within a level may be inhibitory; for example, as one place feature at one time slice is activated it will inhibit the activation of other place features. Connections between units within a level may also be excitatory; for example, a stop consonant at one time slice will facilitate segments that can precede or follow it, such as /s/ or a vowel (depending on the phonotactic constraints of the language).

The excitatory and inhibitory links in TRACE have important implications for the types of processing operations within the model. Because of the inhibitory links within a level, TRACE acts in a 'winner takes all' fashion. Moreover, the of the excitatory links provides a mechanism for the contribution of top-down information to the perception of speech sounds. TRACE contrasts with traditional symbolic invariance approaches because it treats coarticulatory variation as a source of information rather than a source of noise; the inhibitory and facilatory links between one time slice and the next allow for adjacent segments to adjust the weighting to a particular feature or phoneme in a given context (Elman & McClelland, 1986).

Despite these advantages, there are two major problems with TRACE. The first is that although it can use the coarticulatory variation in segmental contexts as information, it is unclear how the model would incorporate other sources of lawful variation such as prosody, rate, or differences among talkers. The second is that TRACE's multiple instantiations of the network across time are considered to be neurally and cognitively implausible (see Cutler, 1995). More recent connectionist models have proposed recurrent neural networks as a way of representing the temporal nature of speech (Elman, 1990; Norris, 1990).

Connectionist models such as TRACE are similar to FLMP because they rely on continuous rather than discrete representations. Continuous activation levels allow for varying degrees of support for competing perceptual hypotheses. Connectionist models also allow for the evaluation and integration of multiple sources of input and rely on general purpose pattern matching schemes with a best fit algorithm. But the connectionist models and the FLMP differ in the degree to which top down influences can affect low level processes. Massaro (198 claims that connectionist models that have top down and bottom up connections are too powerful, predicting both attested and unattested results. Massaro argues that FLMP allows top down *bias* in the perception process while TRACE's two way connections result in top down induced changes in perceptual sensitivity. This is an open issue in need of further research.

The Neighborhood Activation Model

The neighborhood activation model, or NAM, (Luce, 1986; Luce, Pisoni, & Goldinger, 1990) shares with TRACE the notion that words are recognized in the context of other words. A pool of word candidates is activated by acoustic/phonetic input. However, the pool of activated candidates is drawn from the similarity neighborhood of the word (Landauer & Streeter, 1973; Coltheart, Davelaar, Jonasson, & Besner, 1977; Luce, 1986; Andrews, 1989; Luce & Pisoni, 1998). A similarity neighborhood is the set of words that is phonetically similar to the target word. Relevant characteristics of the similarity neighborhood are its density and neighborhood frequency. The density of a word is the number of words in a neighborhood. The neighborhood frequency of a word is the average frequency of words in the neighborhood. There is a strong frequency bias in the model which allows it to deal with apparent top down word frequency effects without positing explicit bi-directional links. Rather than unfolding over time, similarity in NAM is a static property of the entire word. NAM is least developed as a general model of word recognition, as it assumes not only a phonemic level, but word segmentation as well (see Auer & Luce 1997 for a revised version called PARSYN which resolves some of these problems). Moreover NAM has been implemented only for mono-syllabic words. NAM can account for a specific set of lexical similarity effects not treated in other models, and is attractive because it is grounded in a more general categorization model based on the Probability Choice Rule (R.D. Luce, 1959).

SHORTLIST

SHORTLIST (Norris, 1991, 1994) parses a phonemic string into a set of lexical candidates, which compete for recognition. SHORTLIST can be seen as an evolutionary combination of both the TRACE and Marslen-Wilson's Cohort model (Marslen-Wilson & Welsh 1978). A small set (the shortlist) of lexical candidates compete in a TRACE style activation/competition network. The phonemic string is presented gradually to the model, but candidates with early matches to the string have an advantage, due to their early activation, much like the original cohort model. However, as Cutler (1996) notes, SHORTLIST avoids the cognitive implausibility of TRACE's temporal architecture, which effectively duplicates the network at each time slice. SHORTLIST also avoids the cohort model's over-dependence on word initial information. The model takes phonemic information as its input and strictly bottom-up information determines the initial candidate set. The candidate set is determined by comparing whole words but with each strong (i.e., stressed) syllable acting as a potential word onset. This use of prosodic information sets this model apart from others and gives SHORTLIST the ability to parse words from a phrase or utterance represented as a string of phonemes and allophones.

LAFS

Lexical Access from Spectra (Klatt, 1989), or LAFS, is a purely bottom up model of word recognition which compares the frequency spectrum of the incoming signal to stored templates of frequency spectra of words. The stored templates are context-sensitive spectral prototypes derived from subjective experience with the language and consist of all possible diphone (CV and VC) sequences and all cross-word boundaries in the language, resulting in a very large decoding network. Thus, LAFS addresses the problems of contextual variability by precompiling the coarticulatory and word boundary variations into stored representations in an integrated memory system. The model attempts to address interspeaker and rate based variability by using a best fit algorithm to match incoming spectra with stored spectral templates. LAFS fully bypasses the intermediary featural and segmental stages of processing; the perceptual process consists of finding the best match between the incoming spectra and paths through the network.

The advantages of such a strategy are numerous and have been discussed in detail by Klatt (1989). Rather than discarding allophonic and speaker specific detail through reduction to an abstract symbolic representation such as features or segments, the input spectra are retained in full detail. This frees the model from dealing with problems of acoustic invariance across contexts. Since LAFS does not make segmental phonemic level decisions, because it performs recognition at the word level, there is less data reduction than in traditional phonemic based models. More information can be brought to bear on the lexical decision thereby reducing the probability of error and increasing the ability of the system to recover from error (Miller, 1962). Because the stored prototypes are based on subjective learning, there can be local tuning and there is less chance of over-generalization. The perceptual process is explicit being based on a distance/probability metric (Jelinek, 1985) and the scoring strategy is uniform throughout the network.

Despite the power of the approach, there are several problems with the LAFS strategy that Klatt (1989) acknowledges and some that have been raised since then. The most serious is that while LAFS is constructed to accommodate coarticulatory and word-edge variability, it is unlikely that the distance metric is powerful enough to accommodate the full range of variability seen in spoken language. Furthermore, it is nearly impossible to preprocess and store in memory all the sources of variability cited in the variability section above. Finally, much of the stimulus variability in speech comes not in spectra alone but in timing differences as well (Klatt cites the example of variable onset of prenasalization) and LAFS is not built to accommodate much of the temporal nature of speech variation. LAFS is obviously constructed to model a fully developed adult's perception process and contains some developmentally implausible assumptions (but see Jusczyk 1997 for a developmentally oriented adaptation of LAFS called WRAPSA). Its structure involves a priori knowledge about all possible diphones in the language and all cross word boundary combinations; different languages have varying inventories of speech sounds and different constraints on how these sounds can combine within and across words, yet the model depends on these being precompiled for speech perception and word identification to proceed. Cutler (1995) notes that the redundancy inherent in precompiling all word boundaries for every possible word pair separately is psychologically implausible. In addition, recent phonetic research has found different boundary effects at multiple levels of the prosodic hierarchy. Requiring precompiled boundaries at the foot, intonation phrase, and possibly other levels adds to the psychological implausibility of the model. Finally, because the model is explicitly bottom up, it cannot properly model the top down factors like lexical, prosodic, and semantic bias in the lexical decisions.

43

Exemplar Based Models of Word Recognition

Like LAFS, exemplar based models bypass the reduction of the speech signal to featural and segmental units in the word identification process. However, unlike LAFS, exemplar models are instance based rather than relying on precompiled prototypes stored in memory. In exemplar models, there are no abstract categories (whether learned prototypes or innate features and phonemes). Instead, the set of all experienced instances of a category form the basis for the category. The process of categorization therefore involves computing the similarity of the stimulus to every stored instance of every category (e.g., Hintzman, 1986; Nosofsky 1988; Nosofsky, Kruschke, & McKinley, 1992). Although this type of model behaves as if it works on idealized prototype categories (Hintzman, 1986), categorization is a result of computations and the decision process rather than stored prototypes of the stimulus. Exemplar models of perception and memory are fairly widespread in cognitive psychology, but they have only rarely been applied to speech perception and spoken word recognition (for further background on exemplar models in speech perception see Goldinger, 1997).

As discussed at the beginning of this chapter, one of the motivations for proposing that the speech signal is reduced to abstract categories such as phonemes and features has been the widespread belief that memory and processing limitations necessitate data reduction (Haber, 1967). However, more recent empirical data suggest that earlier theorists largely overestimated memory and processing limitations. There is now ample evidence in speech perception and word recognition literature of long term preservation of instance specific details about the acoustic signal. Goldinger (1997) discusses in detail the motivation for exemplar models and cites evidence of episodic memory for such language relevant cases as faces (e.g., Bahrick, Bahrick, & Wittlinger, 1975), physical dynamics (e.g., Cutting & Kozlowski, 1977), modality of presentation (e.g., Stansbury, Rubin, & Linde, 1973), exact wording of sentences (Keenan, McWhinney, & Mayhew, 1977), and talker specific information in spoken words (e.g., Carterette & Barneby, 1975; Hollien, Majewski, & Docherty, 1982; Papcun, Kreiman, & Davis, 1989; Palmeri, Goldinger, & Pisoni, 1993). Taken together this recent evidence has inspired some psycholinguists and speech researchers to reconsider exemplar based approaches to speech perception and spoken word recognition. Although limited, one of the more explicitly implemented models has been proposed by Johnson (1997), which is based on Kruschke's (1992) connectionist ALCOVE model.

In a description of his model, Johnson (1997) discusses several potential problems that an exemplar approach to speech perception must address to be a realistic model of human spoken word recognition. Because most exemplar models have been developed for the perception of static images, these models must be revised and elaborated to take into account the time-varying nature of spoken language. This problem is addressed by considering the role of short term auditory memory in the processing of speech (Baddeley et al., 1998). The incoming signal is sliced into auditory vectors in both the frequency and time domains. As the signal is processed and encoded, the spectral vectors in the short term auditory buffer are matched to all stored vectors and matches are activated adding to previous activation levels thereby representing a veridical short-term memory of the signal (Crowder, 1981). Introducing time in this way permits the modeling of temporal selective attention and segmentation strategies. For example, language specific segmentation strategies such as those that inspired the SHORTLIST model might be modeled by probing the matrix cyclically for boundary associated acoustic events.

A second problem that must be addressed if exemplar models are to be considered cognitively plausible is that of memory limitations. Although there is now a great deal of evidence that much fine detail about specific instances of spoken language is encoded and stored in memory, it is implausible that each experienced auditory pattern is stored at a separate location in the brain or that these instances can be retrieved. Johnson uses ALCOVE (Kruschke, 1992) as the foundation for his implementation of an exemplar model because it uses a covering (vector) map to store exemplars. Locations on the map represent vectors of possible auditory properties (based on known auditory sensitivity). Johnson suggests that vector quantization might be a useful approach. While the storage and matching mechanisms are different from a purely exemplar based model where each instance is stored as a fully separate trace, Johnson's model preserves much of the instance specific information. Only where two instances are identical (from an auditory point of view) at a particular vector does the model collapse information into one representation.

Top down aspects influences on speech perception pose problems for fully bottom up processing models. It might seem that an exemplar model would leave no room for lexical or semantic bias in the decision process. However, usage frequency, recency, and contextual factors can be modeled with base activation levels and attention weights. For example, a high frequency lexical item would have a high base activation level that is directly tied to the frequency of occurrence with a time decay factor (Nosofsky et al., 1992). As syntactic and semantic conditions increase the predictability of a set of words, the base activation rises. Attention weights can be adjusted to model selective attention—the shrinking and expanding of the perceptual space (Nosofsky, 1986) that has frequently been observed. One example of such perceptual distortion is the *perceptual magnet effect* where the perceptual space appears as if it is warped by prototypes (Kuhl, 1991) resulting in decreased sensitivity to changes along a dimension within the range variation of a particular category but increased sensitivity across a category boundary (see however, Lively & Pisoni, 1998).

Finally, Johnson suggests that exemplar models are also capable of incorporating the production-perception link. As a talker produces an utterance, he/she is also hearing the utterance. Therefore, the set of auditory memory traces that are specific to words produced by the talker can be linked to a set of equivalent sensory-motoric exemplars or articulatory plans.

Like the direct perception approach, exemplar based models of perception bring a radical change from past assumptions. Instead of treating stimulus variation in speech as noise to be removed to aid in the perception of abstract units of speech, variability is treated as inherent to the way experiences with speech are stored in memory. Therefore, variation is a source of information that may be used by the perceiver depending on the demands of the listening situation. The appeal of this approach is its ability to account for a very wide variety of speech perception phenomena in a consistent and principled manner. The extensive work on exemplar modeling in other areas of perception means that, unlike much of the traditional work in speech perception, the fundamentals of the approach have been worked out explicitly. However, as the approach is in its infancy in the field of speech perception, it remains to be seen how it performs when tested in a more rigorous fashion across many different environments.

Conclusion

Research on human speech perception has shown that the perceptual process is highly complex in ways beyond our current understanding or theoretical tools. Speech perception relies on both visual and auditory information which are integrated as part of the perceptual process. Near perfect performance is achieved despite an enormous amount of variability both within and across talkers and across a wide variety of different environmental conditions. In the early days of speech research, it was believed that the perceptual process relied on a few invariant characteristics of the segments which differentiated larger linguistic units like words and utterances. While we may yet find higher order relational invariants which are important features for defining the linguistic categories in language, it has already been demonstrated that listeners make use of the lawful variability in the acoustic signal when perceiving speech. Variability cannot be removed, discarded, or normalized away in any psychologically plausible model of speech perception. Some of the approaches discussed above, which rely on more elaborate notions of perceptual categories and long term encoding, incorporate the variability and non-linearity inherent in the speech signal directly into the perceptual process. These new approaches to speech perception treat the speech signal as information rich and use lawful variability and redundant information rather than treating these properties of speech as extraneous noise to be discarded. We believe that these new approaches to the traditional problems of invariance and non-linearity provide a solution to the previously intractable problem of perceptual constancy despite variability in the signal.

References

Abercrombie, D. (1967). *Elements of general phonetics*. Chicago: Aldine.

Ali, L., Gallager, T., Goldstein, J., & Daniloff, R. (1971). Perception of coarticulated nasality. *Journal of the Acoustical Society of America*, **49**, 538-540.

Anderson, D. (1962) *The number and nature of alternatives as an index of intelligibility*. The Ohio State University.

Andrews, S. (1989). Frequency and neighborhood effects on lexical access: activation or search? *Journal of Experimental Psychology: Learning, Memory & Cognition*, **15**, 802-814.

Bahrick, H., Bahrick, P., & Wittlinger, R. (1987). Habituation as a necessary condition for maintenance rehearsal. *Journal of Experimental Psychology: General*, **104**, 54-75.

Bailey, P. J., & Summerfield, Q. (1980). Information in speech: Observations on the perception of [s]-stop clusters. *Journal of Experimental Psychology*, **6**, 536-563.

Beckman, M., & Edwards, J. (1994). Articulatory evidence for differentiating stress categories. In P. A. Keating (Eds.), *Phonological structure and phonetic form: Papers in laboratory phonology III* (pp. 7-33). Cambridge, UK: Cambridge University Press.

Bever, T. G., Lackner, J., & Kirk, R. (1969). The underlying structures of sentences are the primary units of immediate speech processing. *Perception & Psychophysics*, **5**, 191-211.

Black, J. W., & Mason, H. M. (1946). Training for voice communication. *Journal of the Acoustical Society of America*, **18**, 441-445.

Bladon, R. A. W., Henton, C. G., & Pickering, J. B. (1984). Towards an auditory theory of speaker Normalization. *Language and Communication*, **4**, 59-69.

Blumstein, S. E., & Stevens, K. N. (1980). Perceptual invariance and onset spectra for stop consonants in different vowel environments. *Journal of the Acoustical Society of America*, **66**, 1001-1017.

Broadbent, D. E. (1965). Information processing in the nervous system. *Science*, **150**, 475-462.

Browman, C., & Goldstein, L. (1992). Articulatory phonology: An overview. *Phonetica*, **49**, 155-180.

Browman, C. P. (1978). *Tip of the tongue and slip of the ear: Implications for language processing* Ph.D. dissertation. University of California at Los Angeles, Los Angeles, CA.

Browman, C. P., & Goldstein, L. (1989). Articulatory gestures as phonological units. *Phonology*, **6**, 201-251.

Browman, C. P., & Goldstein, L. (1990). Gestural specification using dynamically-defined articulatory structures. *Journal of Phonetics*, **18**, 299-320.

Byrd, D. (1994) *Articulatory timing in English consonant sequences*. Ph. D. dissertation, University of California at Los Angeles, Los Angeles, CA.

Byrd, D. (1996). Influences on articulatory timing in consonant sequences. *Journal of Phonetics*, **24**, 209-224.

Carterette, E., & Barneby, A. (1975). Recognition memory for voices. In A. Cohen & S. Nooteboom (Eds.), *Structure and process in speech perception* (pp. 246-265). New York: Springer-Verlag.

Chiba, T., & M., K. (1941). *The vowel: Its nature and structure*. Tokyo: Kaiseikan.

Chomsky, N., & Halle, M. (1968). *The sound pattern of English*. New York, NY: Harper & Rowe.

Chomsky, N., & Miller, G. A. (1963). Introduction to the formal analysis of natural language. In R. D. Luce, R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (pp. 269-321). New York, NY: Wiley.

Clements, G. N. (1984). Principles of tone assignment in Kikuyu. In G. N. Clements & J. Goldsmith (Eds.), *Autosegmental studies in Bantu tone* Dordrecht: Foris.

Cole, R., & Cooper, W. (1975). The perception of voicing in English affricates and fricatives. *Journal of the Acoustical Society of America*, **58**, 1280-1287.

Cole, R., Rudnicky, A., Zue, V., & Reddy, D. R. (1978). Speech patterns on paper. In R. Cole (Eds.), *Perception and production of fluent speech* Hillside, NJ: Erlbaum.

Collier, R., & t'Hart, J. (1971). The role of intonation in speech perception. In A. Cohen & S. Nooteboom (Eds.), *Structure and process in speech perception* (pp. 107-123). New York: Springer-Verlag.

Coltheart, M., Davelaar, E., Jonasson, J. T., & Besner, D. (1977). Access to the internal lexicon. In S. Dornic (Eds.), *Attention and performance VI* (pp. 535-555). Hillsdale, NJ: Erlbaum.

Cooper, A. (1991). Laryngeal and oral gestures in English /p, t, k/. In *Proceedings of the XIIth International Congress of Phonetic Sciences 2, University of Provence* (pp. 50-53). Aix-en-Provence, France:

Creelman, C. D. (1957). The case of the unknown talker. *Journal of the Acoustical Society of America*, **29**, 655.

Crowder, R. (1981). The role of auditory memory in speech perception and discrimination. In T. Meyers, J. Laver, & J. Anderson (Eds.), *The cognitive representation of speech* (pp. 167-179). New York: North-Holland.

Cummings, K. E., & Clements, M. A. (1995). Analysis of the glottal excitation of emotionally styled and stressed speech. *Journal of the Acoustical Society of America*. **98**, 88-98.

Cutler, A. (1976). Phoneme-monitoring reaction time as a function of preceding intonation contour. *Perception and Psychophysics*, **20**, 55-60.

Cutler, A. (1995). Spoken word recognition and production. In J. L. Miller & P. D. Eimas (Eds.), *Speech, language, and communication* (pp. 97-137). San Diego: Academic Press.

Cutler, A. (1997). The comparative perspective on spoken-language processing. *Speech Communication*, **21**, 3-15.

Cutler, A., Mehler, J., Norris, D. G., & Segui, J. (1986). The syllable's differing role in the segmentation of French and English. *Journal of Memory and Language*, **25**, 385-400.

Cutler, A., Mehler, J., Norris, D. G., & Segui, J. (1992). The mono-lingual nature of speech segmentation by bilinguals. *Cognitive Psychology*, **24**, 381-410.

Cutler, A., & Norris, D. (1988). The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance*, **14**, 381-410.

Cutting, J., & Kozlowski, L. (1977). Recognizing friends by their walk: Gait perception without familiarity cues. *Bulletin of the Psychonomic Society*, **9**, 353-356.

Darwin, C. J. (1975). The dynamic use of prosody in speech perception. In A. Cohen & S. Nooteboom (Eds.), *Structure and process in speech perception* New York: Springer-Verlag.

Darwin, C. J. (1976). The perception of speech. In E.C. Carterette &. M. P. Friedman (Eds.), *Handbook of perception* (pp. 175-216). Heidelberg: Springer-Verlag.

de Jong, K. (1995). The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation. *Journal of the Acoustical Society of America*, **97**, 491-504.

Delattre, P., & Freeman, C. D. (1968). A dialect study of American r's by X-ray motion picture. *Linguistics*, **44**, 29-68.

Delattre, P. C., Liberman, A. M., & Cooper, F. S. (1955). Acoustic loci and transitional cues for consonants. *Journal of the Acoustical Society of America*, **27**(769-773).

Eimas, P. D., & Miller, J. L. (1980). Contextual effects in infant speech perception. *Science*, **209**, 1140-1141.

Eimas, P. D., & Nygaard, L. C. (1992). Contextual coherence and attention in phoneme monitoring. *Journal of Memory and Language*, **31**, 375-395.

Elman, J. L. (1989). Connectionist approaches to acoustic/phonetic processing. In W. D. Marslen-Wilson (Ed.), *Lexical representation and process* (pp. 227-260). Cambridge, MA: MIT Press.

Elman, J. L., & McClelland, J. L. (1986). Exploiting lawful variability in the speech wave. In J. S. Perkell &. D. H. Klatt (Eds.), *Invariance and variability in speech processes* (pp. 360-380). Hillsdale, NJ: Erlbaum.

Fant, G. (1960). *Acoustic theory of speech production*. The Hague: Mouton.

Fant, G. (1962). Descriptive analysis of the acoustic aspects of speech. *Logos*, **5**, 3-17.

Fant, G. (1970). Automatic recognition and speech research. *Speech Transmission Laboratory, Quarterly Progress and Status Report, b, 1/1970*, Royal Institute of Technology, Stockholm.

Flanagan, J. L. (1972). *Speech analysis, synthesis and perception*. New York, NY: Academic Press.

Flege, J. E., & Massey, K. P. (1980). *English prevoicing: random or controlled*. Paper presented at the Linguistic Society of America, 2 August, Albuquerque, NM.

Fletcher, H. (1929). *Speech and hearing*. Princeton, NJ: Von Nostrand Reinhold Company.

Fodor, J. A. (1983). *The modularity of mind*. Cambridge, MA: MIT Press.

Forster, K. I. (1976). Accessing the mental lexicon. In R. J. Wales & E. Walker (Eds.), *New approaches to language mechanisms* Amsterdam: North-Holland.

Fougeron, C., & Keating, P. A. (1996). Variations in velic and lingual articulation depending on prosodic position. *UCLA Working Papers in Phonetics*, **92**, 88-96.

Fougeron, C., & Keating, P. A. (1997). Articulatory strengthening at edges of prosodic domains. *Journal of the Acoustical Society of America*, 3728-3740.

Fowler, C. A. (1980). Coarticulation and theories of extrinsic timing. *Journal of Phonetics*, **8**, 113-133.

Fowler, C. A. (1986). An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics*, **14**, 3-28.

Fowler, C. A. (1995). Speech production. In J. L. Miller & P. D. Eimas (Eds.), *Speech, language, and communication* (pp. 30-62). San Diego: Academic Press.

Fowler, C. A., & Housum, J. (1987). Talkers' signalling of 'new' and 'old' words in speech and listeners' perception and use of the distinction. *Memory and Language*, **26**, 489-504.

Fowler, C. A., & Rosenblum, L. D. (1990). Duplex perception: A comparison of monosyllables and slamming doors. *Journal of Experimental Psychology: Human Perception and Performance*, **16** (4), 742-754.

Fowler, C. A., & Rosenblum, L. D. (1991). The perception of phonetic gestures. In I. Mattingly & M. Studdert-Kennedy (Eds.), *Modularity and the motor theory of speech perception* (pp. 33-59). Hillsdale, NJ: Erlbaum.

Fowler, C. A., & Smith, M. R. (1986). Speech perception as "vector analysis": an approach to the problems of segmentation and invariance. In J. Perkell & D. H. Klatt (Eds.), *Invariance and variability of speech processes* Hillsdale, NJ: Erlbaum.

Frazier, L. (1995). Issues of representation in psycholinguistics. In J. L. Miller & P. D. Eimas (Eds.), *Speech, language, and communication* (pp. 1-29). San Diego: Academic Press.

French, N. R., & Steinberg, J. C. (1947). Factors governing the intelligibility of speech sounds. *Journal of the Acoustical Society of America*, **19**, 90-119.

Fromkin, V. (1965) *Some phonetic specifications of linguistic units: An electromyographic investigation*. Ph. D. dissertation, University of California at Los Angeles, Los Angeles, CA.

Fujimura, O., Macchi, M. J., & Streeter, L. A. (1978). Perception of stop consonants with conflicting transitional cues: A cross-linguistic study. *Language and Speech*, **21**, 337-345.

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology*, **6**, 110-125.

Garnes, S., & Bond, Z. (1980). A slip of the ear: A snip of the ear? A slip of the year? In V. Fromkin (Ed.), *Errors in linguistic performance: Slips of the tongue, ear, pen, hand*. New York, NY: Academic Press.

Gay, T. (1978). Effect of speaking rate on vowel formant movements. *Journal of the Acoustical Society of America*, **63**, 223-230.

Gerstman, L. (1968). Classification of self-normalized vowels. *IEEE Trans. Audio Electroacoust.*, *AU*-**16**, 78-80.

Gibson, J. J. (1966). *The senses considered as perceptual systems*. Boston, MA: Houghton-Mifflin.

Goldinger, S. D. (1990). Neighborhood density effects for high frequency words: Evidence for activation-based models of word recognition. *Research on Speech Perception: Progress Report*, **15**, 163-186.

Goldinger, S. D. (1997). Words and voices: perception and production in an episodic lexicon. In K. Johnson & J. W. Mullennix (Eds.), *Talker variability in speech processing* (pp. 33-66). San Diego: Academic Press.

Goldinger, S. D., Pisoni, D. B., & Logan, J. S. (1991). On the locus of talker variability effects in recall of spoken word lists. *Journal of Experimental Psychology*, **17**, 152-162.

Goldinger, S. D., Pisoni, D. B., & Luce, P. A. (1996). Speech perception and spoken word recognition: Research and theory. In N. J. Lass (Ed.), *Principles in experimental phonetics* (pp. 277-327). St. Louis: Mosby.

Goldsmith, J. A. (1990). *Autosegmental & metrical Phonology*. Oxford, UK: Basil Blackwell.

Halle, M. (1985). Speculations about the representation of words in memory. In V. Fromkin (Ed.), *Phonetic linguistics* (pp. 101-114). New York, NY: Academic Press.

Hagiwara, R. (1995). *Acoustic realizations of American /r/ as produced by women and men*. Ph. D. dissertation, University of California at Los Angeles, Los Angeles.

Harris, K. S. (1958). Cues for the discrimination of American English fricatives in spoken syllables. *Language and Speech*, **1**, 1-7.

Hawkins, S., & Stevens, K. N. (1985). Acoustic and perceptual correlates of the nonnasal-nasal distinction for vowels. *Journal of the Acoustical Society of America*, **77**, 1560-1575.

Heinz, J. M., & Stevens, K. N. (1961). On the properties of voiceless fricative consonants. *Journal of the Acoustical Society of America*, **33**, 589-596.

Henton, C. G. (1988). Creak as a sociophonetic marker. In L. Hyman & C. Li (Eds.), *Language, speech and mind: Studies in honour of Victoria A. Fromkin* (pp. 3-29). London: Routledge.

Henton, C. G., & Bladon, R. A. W. (1985). Breathiness in normal female speech: inefficiency versus desirability. *Language Communication*, **5**, 221-227.

Hintzman, D. L. (1986). Schema Abstraction in a multiple-trace memory model. *Psychological Review*, **93**, 411-423.

Hockett, C. (1955). *Manual of phonology*. Bloomington, IN: Indiana University Press.

Hollien, H., Majewski, W., & Docherty, E. T. (1982). Perceptual identification of voices under normal, stress, and disguise speaking conditions. *Journal of Phonetics*, **10**, 139-148.

House, A. S. (1957). Analog studies of nasal consonants. *Journal of Speech and Hearing Research*, **22**, 190-204.

Hudak, T. J. (1987). Thai. In B. Comrie (Eds.), *The world's major languages* (pp. 757-776). Oxford, UK: Oxford University Press.

Ingemann, F. (1968). Identification of the speaker's sex from voiceless fricatives. *Journal of the Acoustical Society of America*, **44**, 1142-1144.

Jacobson, R., Fant, G., & Halle, M. (1952). *Preliminaries to speech analysis*. Cambridge, MA: M.I.T. Acoustics Laboratory.

Jacoby, L. L., & Brooks, L. R. (1984). Nonanalytic cognition: Memory, perception, and concept learning. In G. Bower (Eds.), *The psychology of learning and motivation* Orlando, FL: Academic Press.

Jelinek, F. (1982). The development of an experimental discrete dictation recognizer. *Proceedings of the IEEE*, **73**, 1616-1624.

Johnson, K. (1997). Speech perception without speaker normalization. In K. Johnson & J. W. Mullennix (Eds.), *Talker variability in speech processing* (pp. 145-165). San Diego: Academic Press.

Johnson, K., Ladefoged, P., & Lindau, M. (1993). Individual differences in vowel production. *Journal of the Acoustical Society of America*, **94**, 701-714.

Joos, M. (1948). Acoustic phonetics. *Language, Suppl.* **24**, 1-136.

Jun, S. A. (1993) *The phonetics and phonology of Korean prosody*. Ph. D dissertation., The Ohio State University, Columbus, OH.

Kewley-Port, D. (1982). Measurement of formant transitions in naturally produced consonant-vowel syllables. *Journal of the Acoustical Society of America*, **73**, 322-335.

Kewley-Port, D., Pisoni, D. B., & Studdert-Kennedy, M. (1983). Perception of static and dynamic acoustic cues to place of articulation in initial stop consonants. *Journal of the Acoustical Society of America*, **73**, 1779-1793.

Kirk, P. J., Ladefoged, J., & Ladefoged, P. (1993). Quantifying acoustic properties of modal, breathy and creaky vowels in Jalapa Mazatec. In A. Mattina & T. Montler (Eds.), *American Indian linguistics and ethnography in honor of Laurence C. Thompson* (pp. 435-450). University of Montana.

Klatt, D. H. (1975). Voice onset time, frication, and aspiration in word-initial consonant clusters. *Journal of Speech and Hearing Research*, **18**, 686-706.

Klatt, D. H. (1976). Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *Journal of the Acoustical Society of America*, **59**, 1208-1221.

Klatt, D. H. (1979). Speech perception: A model of acoustic-phonetic analysis and lexical access. *Journal of Phonetics*, **7**, 279-312.

Klatt, D. H. (1989). Review of selected models of speech perception. In W. D. Marslen-Wilson (Eds.), *Lexical representation and process* (pp. 169-226). Cambridge, MA: MIT Press.

Klatt, D. H., & Cooper, W. E. (1975). Perception of segment duration in sentence contexts. In A. Cohen & S. G. Nooteboom (Eds.), *Structure and process n speech perception* New York, NY: Springer-Verlag.

Klatt, D. H., & Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America*, **87**, 820-857.

Knudsen, V. O. (1929). The hearing of speech in auditoriums. *Journal of the Acoustical Society of America*, **1**, 56-82.

Krakow, R. A. (1989) *The articulatory organization of syllables: A kinematic analysis of labial and velar gestures*. Ph. D dissertation., Yale University, New Haven, CT.

Kruschke, J. K. (1992). ALCOVE: An exemplar based connectionist model of category learning. *Psychological Review*, **99**, 22-44.

Kuehn, D. P., & Moll, K. L. (1973). A cineradiographic study of VC and CV articulatory velocities. *Journal of Phonetics*, **4**, 303-320.

Kuhl, P. K. (1991). Human adults and human infants show a “perceptual magnet effect” for the prototypes of speech categories, monkeys do not. *Perception & Psychophysics*, **50**, 93-107.

Kurowski, K., & Blumstein, S. E. (1984). Perceptual integration of the murmur and formant transitions for place of articulation in nasal consonants. *Journal of the Acoustical Society of America*, **76**, 383-390.

Ladefoged, P. (1968). *Three areas of experimental phonetics*. Oxford, UK: Oxford University Press.

Ladefoged, P., & Broadbent, D. E. (1957). Information conveyed by vowels. *Journal of the Acoustical Society of America*, **29**, 948-104.

Ladefoged, P., & Maddieson, I. (1996). *The sounds of the world's languages*. Oxford, UK: Blackwell.

Ladefoged, P., Maddieson, I., & Jackson, M. T. T. (1988). Investigating phonation types in different languages. In O. Fujimura (Ed.), *Vocal physiology: Voice production, mechanisms and functions* (pp. 297-317). New York: Raven.

Landauer, T. K., & Streeter, L. A. (1973). Structural differences between common and rare words: Failure of equivalence assumptions for theories of word recognition. *Journal of Verbal Learning and Verbal Behavior*, **12**, 119-131.

Lehiste, I. (1970). *Suprasegmentals*. Cambridge: MIT Press.

Lehiste, I. (1976). Role of duration in disambiguating syntactically ambiguous sentences. *Journal of the Acoustical Society of America*, **60**, 1199-1202.

Liberman, A. L. (1957). Some results of research on speech perception. *Journal of the Acoustical Society of America*, **29**, 117-123.

Liberman, A. L., Cooper, F. S., Harris, K. S., & MacNeilage, P. F. (1963). A motor theory of speech perception. In G. Fant (Eds.), *Proceedings of the speech perception seminar, Stockholm, 1962*. Stockholm: Royal Institute of Technology, Speech Transmission Laboratory.

Liberman, A. L., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, **74**, 431-461.

Liberman, A. L., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, **21**, 1-36.

Liberman, A. M., Delattre, P. C., & Cooper, F. S. (1952). The role of selected stimulus-variables in the perception of unvoiced stops. *American Journal of Psychology*, **LVX**, 497-516.

Liberman, A. M., Delattre, P. C., Cooper, F. S., & Gerstman, L. J. (1954). The role of consonant-vowel transitions in the perception of the stop and nasal consonants. *Journal of Experimental Psychology*, **52**, 127-137.

Liberman, A. M., Delattre, P. C., & Gerstman, L. J. (1956). Tempo of frequency change as a cue for distinguishing classes of speech sounds. *Psychological Monogram (Gen. Appl.)*, **68**, 1-13.

Liberman, A. M., & Mattingly, I. G. (1989). A specialization for speech perception. *Science*, **243**, 489-494.

Lieberman, P. (1963). Some effects of semantic and grammatical context on the production and perception of speech. *Language and Speech*, **6**, 172-187.

Lindblom, B. (1963). Spectrographic study of vowel reduction. **35**, 1773-1781.

Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H and H theory. In W. Hardcastle & A. Marchal (Eds.), *Speech production and speech modelling* (pp. 403-439). Dordrecht: Kluwer.

Lindblom, B., & Studdert-Kennedy, M. (1967). On the role of formant transitions in vowel recognition. *Journal of the Acoustical Society of America*, **42**, 830-843.

Lindblom, B., & Svensson, S. G. (1973). Interaction between segmental and nonsegmental factors in speech recognition. *IEEE Transactions on Audio and Electroacoustics, AU-21*, 536-545.

Lisker, L., & Abramson, A. D. (1964). A cross-language study of voicing in initial stops: Acoustic measurements. *Word*, **20**, 384-422.

Lively, S. E., Pisoni, D. B., & Goldinger, S. D. (1994). Spoken word recognition: research and theory. In M. A. Gernsbacher (Eds.), *Handbook of psycholinguistics* (pp. 265-301). New York: Academic Press.

Luce, P. A. (1986) *Neighborhoods of words in the mental lexicon*. Ph. D dissertation., Indiana University, Bloomington, IN.

Luce, R. D. (1959). *Individual choice behavior*. New York: Wiley.

Mack, M., & Blumstein, S. E. (1983). Further evidence of acoustic invariance in speech production: The stop-glide contrast. *Journal of the Acoustical Society of America*, **73**, 1739-1750.

MacNeilage, P. F. (1970). Motor control of serial ordering of speech. *Psychological Review*, 77, 182-196.

Maddieson, I. (1984). *Patterns of sound*. Cambridge: Cambridge University Press.

Maeda, S. (1976) *A characterization of American English intonation*. Ph. D., MIT.

Malécot, A. (1956). Acoustic cues for nasal consonants. *Language*, **32**, 274-278.

Marslen-Wilson, W. D., & Welsh, A. (1978). Processing interactions during word-recognition in continuous speech. *Cognitive Psychology*, **10**, 29-63.

Martin, C. S., Mullennix, J. W., Pisoni, D. B., & Sommers, M. S. (1989). Effects of talker variability on recall of spoken word lists. *Journal of Experimental Psychology: Learning, Memory and Cognition*, **15**, 676-684.

Massaro, D. W. (1987). *Speech perception by ear and eye: A paradigm for psychological inquiry*. Hillsdale, NJ: Erlbaum.

Massaro, D. W. (1989). Multiple book review of speech perception by ear and eye: A paradigm for psychological inquiry. *The Behavioral and Brain Sciences*, **12**, 741-755.

Massaro, D. W., & Cohen, M. M. (1983). Evaluation and integration of visual and auditory information in speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, **9**, 753-771.

McCawley, J. D. (1968). *The phonological component of the Japanese grammar*. The Hague: Mouton.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, **18**, 1-86.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, **264**, 746-748.

Mehler, J. (1981). The role of syllables in speech processing. *Philosophical transactions of the Royal Society of London, Series B*, **295**, 333-352.

Mehler, J., Dommergues, U., Frauenfelder, U., & Segui, J. (1981). The syllable's role in speech segmentation. *Journal of verbal learning and verbal behaviour*, **2020**, 298-305.

Miller, G. A. (1962). Decision units in the perception of speech. *IRE Transactions on Information Theory*, 81-83.

Miller, G. A., Heise, G. A., & Lichten, W. (1951). The intelligibility of speech as a function of the context of the test materials. *Journal of experimental Psychology*, **16**, 329-335.

Miller, G. A., & Nicely, P. E. (1955). An analysis of perceptual confusions among some English consonants. *Journal of the Acoustical Society of America*, **27**, 329-335.

Miller, J. L. (1981). Effects of speaking rate on segmental distinctions. In P. D. Eimas & J. L. Miller (Eds.), *Perspectives on the study of speech* (pp. 39-74). Hillsdale, NJ: Erlbaum.

Miller, J. L. (1987). Rate-dependent processing in speech perception. In A. Ellis (Ed.), *Progress in the psychology of language*. Hillsdale, NJ: Erlbaum.

Miller, J. L. (1990). Speech perception. In D. N. Osherson & H. Lasnik (Eds.), *An invitation to cognitive science* (pp. 69-93). Cambridge, MA: MIT Press.

Miller, J. L., & Baer, T. (1983). Some effects of speaking rate on the production of /b/ and /w/. *Journal of the Acoustical Society of America*, **73**, 1751-1755.

Miller, J. L., Green, K. P., & Reeves, A. (1986). Speaking rate and segments: A look at the relation between speech production and perception for the voicing contrast. *Phonetica*, **43**, 106-115.

Miller, J. L., Grosjean, F., & Lomato, C. (1984). Articulation rate and its variability in spontaneous speech: A reanalysis and some implications. *Phonetica*, **41**, 215-225.

Miller, J. L., & Liberman, A. M. (1979). Some effects of later occurring information on the perception of stop consonant and semivowel. *Perception & Psychophysics*, **25**, 457-465.

Neary, T. M. (1990). The segment as a unit of speech perception. *Journal of Phonetics*, **18**, 347-373.

Neary, T. M. (1992). Context effects in a double-weak theory of speech perception. *Language & Speech*, **35**, 153-172.

Neary, T. M. (1997). Speech perception as pattern recognition. *Journal of the Acoustical Society of America*, **101**(6), 3241-3254.

Nguyen, D. (1987). Vietnamese. In B. Comrie (Eds.), *The world's major languages* (pp. 777-796). Oxford, UK: Oxford University Press.

Norris, D. G. (1991). Rewriting lexical networks on the fly. In *Proceedings of EUROSPEECH 91, Genoa*, **1** (pp. 117-120).

Norris, D. G. (1994). SHORTLIST: A connectionist model of continuous speech recognition. *Cognition*, **52**, 189-234.

Norris, D. G., & Cutler, A. (1988). The relative accessibility of phonemes and syllables. *Perception & Psychophysics*, **43**, 541-550.

Norris, J., & Cutler, A. (1995). Competition and segmentation in spoken-word recognition. *Journal of Experimental Psychology: Learning, Memory and Cognition*, **21**(5), 1209-1228.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, **115**(1), 39-57.

Nosofsky, R. M. (1988). Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **14**, 700-708.

Nosofsky, R. M. (1991). Tests of an exemplar model for relating perceptual classification and recognition memory. *Journal of experimental Psychology: Learning, Memory, and Cognition*, **14**, 3-27.

Nosofsky, R. M., Kruschke, J. K., & McKinley, S. C. (1992). Combining exemplar based category representations and connectionist learning rules. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **18**, 211-233.

Nusbaum, H. C., Schwab, E. C., & Sawusch, J. R. (1983). The role of "chirp" identification in duplex perception. *Perception and Psychophysics*, **33**, 323-332.

Nygaard, L. C., & Pisoni, D. B. (1995). Speech perception: New directions in research and theory. In J. L. Miller & P. D. Eimas (Eds.), *Speech, language, and communication* (pp. 63-96). San Diego: Academic Press.

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1995). Effects of stimulus variability on perception and representation of spoken words in memory. *Perception & Psychophysics*, **57**, 989-1001.

O'Connor, J. D., Gerstman, L. J., Liberman, A. M., Delattre, P. C., & Cooper, F. S. (1957). Acoustic cues for the perception of initial /w, j, r, l/ in English. *Word*, **13**, 22-43.

Oden, G. C., & Massaro, D. W. (1978). Integration of featural information in speech perception. *Psychological Review*, **85**, 172-191.

Palmeri, T. J., Goldinger, S. D., & Pisoni, D. B. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology: Learning, Memory and Cognition*, **19**, 1-20.

Papcun, G., Kreiman, J., & Davis, A. (1989). Long-term memory for unfamiliar voices. *Journal of the Acoustical Society of America*, **85**, 913-925.

Pastore, R. E., Schmuckler, M. A., Rosenblum, L., & Szczesiul, R. (1983). Duplex perception with musical stimuli. *Perception and Psychophysics*, **33**, 469-474.

Peters, R. W. (1955). The relative intelligibility of single-voice messages under various conditions of noise. *Joint Report No. 56, U.S. Naval School of Aviation Medicine* (pp. 1-9). Pensacola, FL.

Pierrehumbert, J., & Beckman, M. (1988). *Japanese tone structure*. Cambridge, MA: MIT Press.

Pisoni, D. B. (1977). Identification and discrimination of the relative onset of two component tones: Implications for voicing perception in stops. *Journal of the Acoustical Society of America*, **61**, 1352-1361.

Pisoni, D. B. (1978). Speech perception. In W. K. Estes (Eds.), *Handbook of learning and cognitive processes* (pp. 167-233). Hillsdale, NJ: Erlbaum.

Pisoni, D. B. (1990). Effects of talker variability on speech perception: Implications for current research and theory. In *Proceedings of the 1992 International Conference on Spoken Language Processing* (pp. 587-590). Banff, Canada:

Pisoni, D. B. (1997). Some thoughts on "normalization" in speech perception. In K. Johnson & J. W. Mullennix (Eds.), *Talker variability in speech processing* (pp. 9-32). San Diego: Academic Press.

Pollack, I., & Pickett, J. M. (1964). The intelligibility of excerpts from conversations. *Language and Speech*, **6**, 165-171.

Rand, T. C. (1974). Dichotic release from masking for speech. *Journal of the Acoustical Society of America*, **55**, 678-680.

Remez, R. E. (1986). Realism, language, and another barrier. *Journal of Phonetics*, **14**, 89-97.

Remez, R. E. (1987). Units of organization and analysis in the perception of speech. In M. E. H. Schouten (Eds.), *The psychophysics of speech perception* (pp. 419-432). Dordrecht: Martinus Nijhoff.

Remez, R. E., Rubin, P. E., Berns, S. M., Pardo, J. S., & Lang, J. M. (1994). On the perceptual organization of speech. *Psychological Review*, **101**(1), 129-156.

Repp, B. H. (1982). Phonetic trading relations and contest effects: New experimental evidence for a speech mode of perception. *Psychological Bulletin*, **92**, 81-110.

Repp, B. H. (1983a). Categorical perception: Issues, methods, findings. *Speech and Language: Advances in Basic Research and Practice*, **10**.

Repp, B. H. (1983b). Bidirectional contrast effects in the perception of VC-CV sequences. *Perception and Psychophysics*, **33**, 147-155.

Repp, B. H. (1979). Relative amplitude of aspiration noise as a voicing cue for syllable-initial stop consonants. *Language and Speech*, **22**, 173-189.

Repp, B. H., & Mann, V. A. (1981). Perceptual assessment of fricative-stop coarticulation. *Journal of the Acoustical Society of America*, **69**, 1154-1163.

Salasoo, A., & Pisoni, D. B. (1985). Interaction of knowledge sources in spoken word identification. *Journal of Memory and Language*, **24**, 210-231.

Samuel, A. G. (1981). The role of bottom-Up Confirmation in the Phonemic Restoration Illusion. *Journal of Experimental Psychology: Human Perception and Performance*, **7**, 1124-1131.

Samuel, A. G. (1982). Phonetic prototypes. *Perception & Psychophysics*, **31**, 307-314.

Schadle, C. (1985) *The acoustics of fricative consonants*. Cambridge, MA: MIT Press.

Schwartz, M. F. (1968). Identification of speaker sex from isolated, voiceless fricatives. *Journal of the Acoustical Society of America*, **43**, 1178-1179.

Segui, J. (1984). The syllable: A basic perceptual unit in speech processing. In H. Bouma & D. G. Bouwhuis (Eds.), *Attention and performance X: Control of language processes* Hillsdale, NJ: Erlbaum.

Shinn, P., & Blumstein, S. E. (1984). On the role of the amplitude envelope for the perception of [b] and [w]. *Journal of the Acoustical Society of America*, **75**, 1243-1252.

Silverman, D. (1997). Pitch discrimination during breathy versus modal phonation (final results). *Journal of the Acoustical Society of America*, **102**(5), 3204.

Skinner, T. E. (1977). Speaker invariant characteristics of vowels, liquids, and glides using relative formant frequencies. *Journal of the Acoustical Society of America*, **62**(S1), S5.

Soli, S. D. (1982). Structure and duration of vowels together specify fricative voicing. *Journal of the Acoustical Society of America*, **72**, 366-378.

Sommers, M. S., Nygaard, L. C., & Pisoni, D. B. (1992). The effects of speaking rate and amplitude variability on perceptual identification. *Journal of the Acoustical Society of America*, **91**, 2340.

Steinberg, J. C. (1929). Effects of distortion on telephone quality. *Journal of the Acoustical Society of America*, **1**, 121-137.

Steriade, D. (1993). Closure release and nasal contours. In R. Krakow & M. Huffman (Eds.), *Nasals, nasalization, and the velum* (pp. 401-470). San Diego: Academic Press.

Stevens, K. N. (1972). Sources of inter- and intra- speaker variability in the acoustic properties of speech sounds. In A. Rigault & R. Charbonneau (Eds.), *Proceedings of the 7th International Congress of Phonetic Sciences* (pp. 206-232). The Hague: Mouton.

Stevens, K. N., & Blumstein, S. E. (1981). The search for invariant acoustic correlates of phonetic features. In P. D. Eimas & J. L. Miller (Eds.), *Perspectives on the study of speech* (pp. 1-38). Hillsdale, NJ: Erlbaum.

Stevens, K. N., & House, A. S. (1955). Development of a quantitative description of vowel articulation. *Journal of the Acoustical Society of America*, **27**, 484-493.

Stevens, K. N., & House, A. S. (1963). Perturbation of vowel articulations by consonantal context: An acoustic study. *Journal of Speech and Hearing Research*, **6**, 111-128.

Strange, W., Jenkins, J. J., & Johnson, T. L. (1983). Dynamic specification of coarticulated vowels. *Journal of the Acoustical Society of America*, **74**, 695-705.

Studdert-Kennedy, M. (1974). The perception of speech. In T. A. Sebeok (Eds.), *Current trends in linguistics* (pp. 2349-2385). The Hague: Mouton.

Studdert-Kennedy, M. (1976). Speech perception. In N. J. Lass (Eds.), *Contemporary issues in experimental linguistics* (pp. 213-293). New York: Academic Press.

Studdert-Kennedy, M. (1980). Speech perception. *Language and Speech*, **23**, 45-65.

Studdert-Kennedy, M. (1982). On the dissociation of auditory and phonetic perception. In R. Carlson & B. Granstrom (Eds.), *The representation of speech in the peripheral auditory system*. Elsevier Biomedical Press.

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, **26**, 212-215.

Summerfield, Q. (1975). *Acoustic and phonetic components of the influence of voice changes and identification times for CVC syllables*. Report of Speech Research in Progress, 4. Queen's University of Belfast.

Summerfield, Q. (1981). On articulatory rate and perceptual constancy in phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance*, 7, 1074-1095.

Summerfield, Q., & Haggard, M. P. (1973). *Vocal tract normalisation as demonstrated by reaction time*. Report of Speech Research in Progress, 2. Queen's University of Belfast.

Sussman, H. M. (1988). The neurogenesis of phonology. In H. A. Whitaker (Ed.), *Phonological processes and brain mechanisms* (pp. 1-23). New York: Springer-Verlag.

Sussman, H. M., McCaffrey, H. A., & Matthews, S. A. (1991). An investigation of locus equations as a source of relational invariance for stop place categorization. *Journal of the Acoustical Society of America*, **90**, 1256-1268.

Sussman, H. R. (1989). Neural coding of relational invariance in speech: Human language analog to the barn owl. *Psychological Review*, **96**, 631-642.

Syrdal, A. K., & Gopal, H. S. (1986). A perceptual model of vowel recognition based on the auditory representation of American English vowels. *Journal of the Acoustical Society of America*, **79**, 1086-1100.

Townsend, J. T. (1989). Winning "20 questions" with mathematical models. *The Behavioral and Brain Sciences*, **12**, 775-776.

Treisman, M. (1971). On the word frequency effect: Comments on the papers by J. Catlin and L. H. Nakatani. *Psychological Review*, **17**, 37-59.

Varya, M., & Fowler, C. A. (1992). Declination of supralaryngeal gestures in spoken Italian. *Phonetica*, **49**, 48-60.

Vassière, J. (1986). Comment on Abbs's paper. In J. S. Perkell & D. H. Klatt (Eds.), *Invariance and variability in speech processes* (pp. 1205-1216). Hillsdale, NJ: Erlbaum.

Vassière, J. (1988). Prediction of velum movement from phonological specifications. *Phonetica*, **45**, 122-139.

Walley, A., & Carrell, T. (1983). Onset spectra and formant transitions in the adult's and child's perception of place of articulation in initial stop consonants. *Journal of the Acoustical Society of America*, **73**, 1011-1022.

Warren, R. M. (1970). Perceptual restoration of missing speech sounds. *Science*, **176**, 392-393.

Whalen, D. H., & Liberman, A. M. (1987). Speech perception takes precedence over non-speech perception. *Science*, **237**, 169-171.

Weiringen, A. V. (1995) *Perceiving dynamic speechlike sounds*. Ph.D. dissertation, University of Amsterdam.

Wright, R. (1996) *Consonant clusters and cue preservation in Tsou*. Ph.D. dissertation, University of California at Los Angeles, Los Angeles, CA.

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, **8**, 338-353.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 21 (1996-1997)
Indiana University

Looking at the "Stars": A First Report on the Intercorrelations Among Measures of Speech Perception, Intelligibility and Language Development in Pediatric Cochlear Implant Users[1]

David B. Pisoni,[2] Mario A. Svirsky,[3] Karen Iler Kirk,[3] and Richard T. Miyamoto[3]

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405

---

[1] This work was upported by NIH NIDCD Research Grants DC00064, DC00423 and DC00111 to Indiana University. Revised and expanded paper presented at the Vth International Cochlear Implant Conference, May 1-3, 1997, New York, NY.

[2] Also DeVault Otologic Research Laboratory, Department of Otoloryngology-Head & Neck Surgery, Indiana University School of Medicine, Indianapolis, IN.

[3] DeVault Otologic Research Laboratory, Department of Otoloryngology-Head & Neck Surgery, Indiana University School of Medicine, Indianapolis, IN..

D.B. Pisoni, M.A. Svirsky, K.I. Kirk and R.T. Miyamoto

Looking at the "Stars": A First Report on the Intercorrelations Among Measures of Speech Perception, Intelligibility and Language Development in Pediatric Cochlear Implant Users

Abstract. It is now well-established in the field of pediatric cochlear implantation that some prelingually deaf children perform significantly better on standardized tests of speech perception and spoken word recognition than other prelingually deaf children. The differences are seen most clearly on very difficult open-set tests of spoken word recognition such as the PBK test. The children who do well on this particular test are frequently referred to as the "Stars," and their extraordinarily good performance is typically reported at scientific meetings and highlighted in journal publications. Unfortunately, very little is actually known about the basis for the superior performance of these children or the audiological and psychological factors that predict which children will become "Stars" and which ones will perform closer to the mean. In this paper, we report the initial results of a correlational analysis of a small group of exceptionally good cochlear implant users, the so-called "Stars." Our goal was to identify the primary factors that underlie their exceptionally good performance on open-set tests of speech perception and spoken word recognition. Speech perception, intelligibility and language scores were examined for a group of children who scored in the top 20% on the PBK test two years post-implant and a "control" group of children who scored in the bottom 20% on the PBK test. Separate correlational analyses were then carried out on the test scores for these subjects 1 year post-implant in order to examine the relations among these dependent measures. The results of our analyses revealed that the "Stars" not only display superior performance on the criterial PBK test but also show very high levels of performance on several other speech perception and language tests. Most notable were the unusually high correlations among several different measures of spoken word recognition and scores on the Reynell receptive and expressive language scales. This suggests that one common underlying factor in these children may be the acquisition of language, specifically, the development of the lexicon, which serves as the "interface" between the initial sensory input and the phonological representation of the sound patterns of words in lexical memory. In addition, word recognition performance was also highly correlated with speech intelligibility scores and open-set measures of language comprehension that required children to interpret spoken language in meaningful ways. Taken together, our analyses of these exceptionally good cochlear implant users suggest that the "Stars" are the children who have been able to begin the normal process of language acquisition. The pattern of intercorrelations among several measures of speech perception, intelligibility, and language development suggests that these children are receiving sufficient sound input through their cochlear implants to map sound patterns onto meanings and build a lexicon of words, two necessary prerequisites for constructing a grammar of the target language from the ambient language spoken in their environment.

Introduction

Although there is now substantial clinical evidence demonstrating the efficacy of cochlear implants in deaf children, many important research questions remain to be addressed regarding the effectiveness of cochlear implants in facilitating speech perception and language development in prelingually deafened children. One of the most important problems concerns the enormous individual variability observed among users of cochlear implants (Staller, Pelter, Brimacombe, Mecklenberg, & Arndt, 1991). In both adults and children, some users of cochlear implants do very well on standardized audiological tests that assess speech perception and word recognition skills, whereas many other users do less well, despite comparable psychophysical data on the electrical stimulation provided by the implant. This is a consistent finding reported at every pediatric cochlear implant center (Fryauf-Bertschy, Tyler, Kelsay & Gantz, 1992). Moreover, in prelingually deafened children using cochlear implants, a small number go on to develop highly intelligible speech (Svirsky, 1996; Robbins, Svirsky & Kirk, in press) and show strong evidence of acquiring the grammar of spoken language, that is, a phonology, syntax and semantics. Other prelingually deafened children with cochlear implants never achieve these important milestones in speech and language development although they do derive some benefit from their cochlear implant in terms of improved awareness and recognition of environmental sounds. The primary motivation for a cochlear implant is, of course, to provide deaf children with a means of perceiving speech and acquiring spoken language.

Much of the clinical research on the efficacy of cochlear implants in young children has concentrated on the study of demographic variables in the hopes of predicting outcome performance and guiding decision making regarding education (Miyamoto, Osberger, Todd et al., 1994). We now know from research findings reported in the published literature that a small number of key demographic variables play an important role in predicting the speech perception performance of prelingually deafened children (Osberger, Robbins, Todd, Riley, Kirk & Carney, 1996). Children who receive their implants at an early age, experience relatively short periods of sensory deprivation (Fryauf-Bertschy, Tyler, Kelsay, Gantz & Woodworth, 1997; Osberger, Todd, Berry, Robbins, & Miyamoto, 1991; Staller, Pelter et al., 1991) and are placed in "oral-only" programs often display very high levels of performance on standardized tests of speech perception and word recognition (Miyamoto, Osberger, Robbins, Myres, & Kessler, 1993; Waltzman, Cohen, Gomolin, Shapiro, Ozdaman & Hoffman, 1994). Many of these children subsequently proceed to achieve high scores on a variety of language measures that assess both receptive and expressive linguistic abilities (Kirk, Pisoni & Miyamoto, in press; Robbins, Svirsky & Kirk, in press). Recent analyses of language development (Svirsky, 1996; Miyamoto, Svirsky & Robbins, 1997; Robbins et al., in press) from our laboratory suggest that in some cases, deaf children with cochlear implants appear to be on the same trajectory of developmental change as normal-hearing children, although the implanted children are delayed in time because of their later start on the task of language learning. Follow-up studies suggest that these implanted children continue to improve their language skills, and there is every indication that they will acquire the grammar of the ambient language and eventually catch up to their normal-hearing peers (Robbins, Svirsky & Kirk, in press). These children show substantial gains in speech perception, phonology, syntax and intelligibility with a cochlear implant above and beyond what would be expected developmentally from maturation alone (Svirsky, 1996).

Despite these very encouraging findings on speech perception and language development in some deaf children, it is also well documented in the literature that many other deaf children who receive cochlear implants at an early age do not acquire these same kinds of linguistic skills and abilities (Fryauf-Bertschy et al., 1997). The individual subject variability in the pediatric population is very substantial and although some children are classified as "Stars" because of their extraordinarily good performance on standardized audiological tests, many other children continue to struggle and have difficulty using their cochlear implants to perceive speech and understand spoken language. These children never really manage to produce highly intelligible speech or to acquire the linguistic skills and underlying linguistic competence of normal-hearing children.

At the present time, we do not know the reasons for these differences in performance between the "Stars" and the average or poor users. The finding that some implanted children do display exceptionally good performance can be taken, at first glance, as an "existence proof" for the efficacy of cochlear implants in deaf children: cochlear implants work with some children and can facilitate the processes of speech perception and language development. A major problem now, however, is that cochlear implants do not work well with all children, and some children derive very little, if any, benefit from their implants. Why does this occur? What sensory, perceptual and cognitive factors are responsible for the differences in performance among deaf children with cochlear implants? These are the questions that motivated the present investigation and encouraged us to examine closely individual differences among pediatric cochlear implant users.

In this report, we focus our attention on the "Star" performers, that is, those children who are exceptionally good users of cochlear implants. The question we explored was how these deaf children manage to do so well with their cochlear implant. What sensory, perceptual and cognitive skills do these exceptionally good users have that the other children lack? Learning more about how the "Stars" perceive speech and develop spoken language may be very useful in helping poorer-performing children reach their potential and more generally increasing the effectiveness of cochlear implants in children. If we can identify the factor or set of factors that underlie these individual differences in performance, we might be in a much better position to recommend and introduce changes in the child's post-implant language-learning environment and modify the specific intervention methods used in aural rehabilitation and language therapy. At the present time, without knowing how or why children differ in their speech perception and language skills, clinicians have no principled theoretical basis for recommending or selecting one communication method over another (i.e., oral vs. total communication) for a given child or adopting one procedure for aural rehabilitation and language therapy over another.

An examination of the published literature on cochlear implants in children reveals that very little research has focused on the study of individual differences or a detailed examination of the "Stars." To our knowledge, no one has tried to identify or describe the underlying sensory, perceptual and cognitive abilities of these unusually high-performing children. In previous studies, the criterion used to identify the "Stars" was exceptionally good performance on one particular perceptual test, the Phonetically Balanced Kindergarten (PBK): Words test (Haskins, 1949), which is an open-set word recognition test. Among clinicians, this test is considered to be a very difficult for a prelingually deaf child compared to other, closed-set perceptual tests that are routinely included in a standard assessment battery (Zwolan, Zimmerman-Phillips, Asbaugh, Hieber, Kileny & Telian, 1997). The children who do very well on the PBK test frequently display ceiling levels of performance on many of the other closed-set speech perception tests that measure speech pattern discrimination. In contrast, open-set tests like the PBK measure word recognition and lexical access and require the child to search and retrieve representation of words from lexical memory. These kinds of word recognition tests are extremely difficult because the listener must perceive and encode fine phonetic differences based on information present in the speech signal without the aid of any external context or retrieval cues and then discriminate/select a pattern from a large number of equivalence classes in memory (see Luce & Pisoni, 1998).

To learn more about the "Stars," we analyzed data from pediatric cochlear implant users who scored exceptionally well on the PBK test two years post-implant. The PBK score was used as a criterial variable to identify and select two groups of subjects for subsequent analysis using an extreme groups design. After these subjects were selected, we examined their performance on a variety of measures already obtained from these children as part of a large-scale longitudinal study at Indiana University Medical Center. These scores included standard demographic variables, open- and closed-set speech perception measures, performance on another test of spoken word recognition and lexical access, vocabulary scores, as well as expressive and receptive language measures. Comprehension scores and a measure of speech intelligibility were also examined. The analyses of most interest were the intercorrelations among these different tests. Does a child who performs exceptionally well on the PBK test also perform equally well on other tests of speech perception, word recognition and language? What is the relationship, if any, between performance on the PBK test and speech intelligibility in these children? Is it possible to identify a common underlying variable or factor that can explain the patterns of relationships observed among these measures? Can we identify specific characteristics of the "Stars" that can be used to account for their exceptionally good performance? Answers to these questions will provide important new knowledge and detailed information about the "Stars" and why these children do so exceptionally well on open-set tests of speech perception like the PBK. The findings from the "Stars" may also have some clinical implications for those children who do not do as well with their cochlear implant and suggest how we can improve their performance in these areas of speech perception and language processing.

If the "Stars" are simply better on every measure of performance compared to a control group of implanted children who score very low on the PBK test, this pattern of results would not be very interesting or theoretically important. This particular outcome would simply reflect an overall difference in performance levels and would not provide any new information about the locus or underlying source of the differences in performance between the two groups. In contrast, if we do observe a pattern of selective differences in performance between the two groups on some measures but not others, this outcome would be very informative about the nature of the individual differences and would help to identify where in the information processing system the differences could be located. Identifying selective differences across a range of measures of performance in these children may also help us to determine if the observed differences are due simply to some global sensory, perceptual or cognitive abilities or whether the differences reflect the operation of domain-specific information processing operations that are specific to perceiving speech signals, encoding linguistic information and processing spoken language. Does exceptionally good performance on the PBK test simply reflect a set of isolated word recognition skills or does it indicate some potentially more important underlying set of linguistic skills or abilities that are also present in other measures as well? The answers to these questions have broad theoretical and clinical implications for how we approach the study of individual differences among pediatric cochlear implant users in the future and what specific changes we may recommend for intervention and therapy in the post-implant language learning environment.

Methods

Subjects

This study used an extreme groups design. Two groups of pediatric cochlear implant users, "Stars" and "Controls," were selected from a large longitudinal study providing data from 160 deaf children. Subjects in both groups were all prelingually deafened (mean =0.4 years onset), and they received a cochlear implant because they were unable to derive any benefit from conventional hearing aids. The criterion used to identify the "Stars" and assign subjects to these two groups was based entirely on word recognition scores from the PBK test, a very difficult open-set word recognition test. The "Star" performers were 27 children who scored in the upper 20% on the PBK test two years post-implant. The "Control" subjects were 23 children who scored in the lower 20% on the PBK test two years post-implant. The mean percentage of words correctly recognized on the PBK test was 25.64 for the "Stars" and 0.0 for the "Control" subjects. A summary of the demographic characteristics of the two groups is shown in Table 1. No attempt was made to match the subjects on any other demographic variable. As a result of this selection procedure, the two groups were comparable in terms of age at onset of deafness and length of implant use. As shown in Table I, the two groups did differ in chronological age, age at implantation, length of deprivation and communication mode.

TABLE I

SUMMARY OF DEMOGRAPHIC INFORMATION

	Stars (N = 27)	*Controls* (N = 23)
Mean Age at Onset (Years)	.3	.8
Mean Age at Implantation (FIT) (Years)	5.8	4.4
Mean Length of Deprivation (Years)	5.5	3.6
Mean Chronological Age (Years)	6.7	5.4
Mean Length of Implant Use (Years)	.9	1.0
Communication Mode:		
Oral Communication	N=19	N=8
Total Communication	N=8	N=15

Stimulas Materials

Scores on a variety of behavioral tests routinely used to assess performance on speech perception, spoken word recognition, language development and speech intelligibility were obtained for each child from our database at testing intervals of 1, 2 and 3 years post-implant. In addition, scores on several psychological tests that were originally used as clinical data for determining implant candidacy were also obtained for each subject. These measures included non-verbal intelligence test (P-IQ) scores, results on the VMI (Beery, 1989), a test of visual-motor coordination, as well as several measures of visual attention and vigilance taken from the Gordon Diagnostic System (Gordon, 1987), a continuous performance test used to diagnose ADHA (American Psychiatric Association, 1994).

Speech Perception Tests. Scores were obtained from two speech perception tests, the Minimal Pairs Test (Robbins et al., 1988) and the Common Phrases Test (Osberger et al., 1991). The Minimal Pairs Test (MPT) is a closed-set test that assesses discrimination of manner, voicing and place differences between pairs of words. The Common Phrases Test (CPT) is an open-set test of listening comprehension that requires the child to answer a series of simple questions or follow a command. The Minimal Pairs Test measures speech feature discrimination in isolated words, whereas the Common Phrases Test measures speech perception and spoken language comprehension in meaningful sentences. The MPT assesses a child's ability to discriminate among pairs of words differing by a single vowel or consonant feature using audition alone. In contrast, the CPT assesses speech perception and spoken language comprehension in meaningful sentences in auditory-only, visual-only (lip-reading) and auditory-plus-visual modalities.

Word Recognition Tests. In addition to the PBK test, which was used as a criterial variable to assign subjects to extreme groups, four scores from the Lexical Neighborhood Test (LNT; Kirk, Pisoni & Osberger, 1995) were also obtained for each subject when available. Like the PBK, the LNT is also an open-set word recognition test administered in the auditory-only modality. However, this new word recognition test was designed specifically to assess the degree to which children can make fine phonetic discriminations among acoustically similar patterns when they are required to retrieve spoken words from lexical memory (Kirk, et al., 1995). The LNT has been shown to provide useful diagnostic information about how young children perceive spoken words in isolation and how they organize and access words from lexical memory.

Vocabulary Tests. Scores on a standardized vocabulary test, the Peabody Picture Vocabulary Test-Revised (PPVT; Dunn, 1965), were also obtained for each child. The PPVT is a closed-set test that requires the child to select a spoken word from four pictures. During testing, children who use total communication are presented each target word using simultaneous spoken and signed English. Children who use oral/aural communication are presented with spoken and written forms of the target word. A receptive language age and a language quotient (chronological age divided by receptive language age) are derived for each child.

Language Tests. Scores on a standardized language assessment test were obtained for each child. Measures of both receptive and expressive language development were obtained from the Reynell Developmental Language Scales (Reynell & Gruber, 1990), which has been used extensively to assess language development in deaf children. The test format involves object manipulation and description based on questions and instructions of varied length and grammatical complexity. This measure yields a receptive and expressive language age and a language quotient using normative data obtained from hearing children.

Speech Intelligibility Scores. Speech intelligibility was assessed by obtaining samples of speech using materials from the BIT (Osberger, Robbins, Todd & Riley, 1994) and the Monsen sentences (Monsen, 1983). These utterances were then transcribed by three naïve adult listeners who were unfamiliar with deaf speech. Each subject produced 10 sentences that were repeated after an examiner's spoken model. Children under six years were administered one list from the Beginner's Intelligibility Test (BIT). The BIT utilizes objects and pictures to convey the target sentence, and an imitative response is elicited. Older children ($\geq$ 6 years) who could read were given the Monsen Sentences Test. For this procedure, children read the sentences, and then imitated the examiner's spoken model. Each subjects' sentence productions were audiorecorded, digitized, randomized, and then played to panels of listeners who transcribed what they thought the subject had said. Panels consisted of three listeners with no prior experience in listening to the speech of persons with hearing loss. Each set of 10 sentences produced by one subject was evaluated by a single panel. The listeners evaluated more than one set of sentences but each set was produced by a new subject, and contained a sentence list not previously heard by the panel. Individual subject intelligibility scores were calculated by averaging the number of words correctly transcribed across the three listeners.

Functional Assessments. Scores on two functional assessments, the Meaningful Auditory Integration Scale (MAIS; Robbins, 1990) and the Meaningful Use of Speech Scale (MUSS; Robbins & Osberger, 1990), measured the use and consistency of a child's speech perception and speech production skills in everyday situations from parent/teacher observations. Both instruments use a rating scale format to record responses. The MAIS assesses the meaningful use of sound perception, whereas the MUSS measures the meaningful use of speech and particular speech production skills in different environments.

Procedures

Scores on the speech perception and language tests were obtained for each child at four sampling intervals: (1) Pre-implant, (2) 1-year, (3) 2-years, and (4) 3-years post-implantation. Because the psychological measures were originally used only for clinical purposes in determining implant candidacy and were not a part of the standard research protocol administered at regular testing intervals to all children in the larger longitudinal project, we examined only one set of scores obtained close to the time of implantation. Another separate study is currently underway that examines the changes in these scores as a function of implant use and assesses the correlations of these psychological measures with speech perception, language development and speech intelligibility scores for these children.

Results and Discussion

Demographics

A summary of measures on the major demographic variables is presented in Table I. As shown here, several differences emerged between the two groups which were the result of the selection procedures used to identify the "Stars" and the control subjects based on the PBK scores as a criterial variable. Although small and statistically significant differences were present in chronological age, age at implantation, and length of deprivation between the groups, both groups of subjects were equivalent in age at onset of deafness and length of implant use.

Differences also occurred in the distribution of subjects assigned to each group based on their communication mode. More oral subjects were assigned to the "Stars" and more total communication (TC) subjects were assigned to the control group. Because communication mode is known to produce substantial effects on many outcome measures, we initially included this variable as an additional factor in each analysis of the measures described below. A series of 2 X 2 analyses of variance were initially used to examine main effects and assess interactions of communication mode with each of the dependent variables described below. Except for only a few of the measures noted below, communication mode alone produced only small main effects or interactions that were not statistically reliable. Therefore, unless otherwise indicated, we present the combined data in the analyses reported below.

Pre-implant scores on the performance measures described below were not considered here because they were generally quite low for all subjects and our major interest in this report is the pattern of correlations among measures of speech perception and language that emerge after implantation. These findings, along with the psychological data obtained on IQ, VMI and visual attention tests will be reported in another paper that examines longitudinal changes in these measures with implant use and their intercorrelations.

Group Differences: Stars vs. Controls

In this section, we present a summary of the group differences obtained for each of the dependent measures. These results provide an initial benchmark for all subsequent comparisons between "Stars" and "Controls" and for the correlational analyses reported below.

Minimal Pairs Test. A summary of the results from the Minimal Pairs Test (MPT) is shown for both groups in Figure 1. Percent correct discrimination scores are presented separately for manner, voicing and place contrasts as a function of implant use in years. Data for the "Stars" are shown by the filled dark bars; data for the "Controls" are shown by the light bars. The Minimal Pairs Test is a two-alternative forced-choice speech feature discrimination task with chance performance at 50 percent correct. This level of performance is indicated in Figure 1 by a dotted horizontal line. A second dotted line is also shown in this figure representing discrimination scores that are significantly above chance performance using the binomial distribution.

Insert Figure 1 about here.

The MPT discrimination data in Figure 1 show several consistent findings. First, performance of the "Stars" on this test is consistently better than the control subjects for every comparison shown in the figure. Second, discrimination performance improves over time with implant use for both groups, although most of the data points in the figure, especially for the control subjects, are not statistically above chance expectation. Finally, discrimination of manner differences by the "Stars" is reliably above chance expectation for all three years and is better overall than voicing and place discrimination which show performance levels above chance only after three years of implant use. Although there are increases in performance over time on all three contrasts for the control subjects, performance never reaches the levels observed with the "Stars" in any of the conditions nor does it deviate significantly from chance expectation, although performance is consistently above chance for almost all of the comparisons. A 3 x 3 x 2 factorial analysis of variance was carried out to assess these differences. The three main effects, group, use and speech feature were all highly significant ($p<.0001$). In addition, there was a significant interaction between group and speech feature ($p<.0005$). Manner was discriminated better than voicing and place, which were not significantly different from each other. Although performance on the MPT for the "Stars" exceeded chance for manner, voicing and place after three years of implant use, performance for the "Controls" was never greater than chance levels even after three years.

Common Phrases Test. The results of the Common Phrases Test are shown in Figure 2 for both the "Stars" and "Controls" as a function of length of implant use for the three presentation modes of this test. Common Phrases is an open-set comprehension test so that chance performance is theoretically zero. This test is routinely presented in three formats—auditory-only (CPA), visual-only (CPV) and combined auditory plus visual (CPAV). Performance levels of the "Stars" on all three formats were consistently better than the "Control" group at each testing interval. And, both groups showed increases in performance as a function of length of implant use. The multi-modal presentation condition (CPAV) was better than either the auditory-only or visual-only, which also differed reliably from each other. Analysis of variance revealed that all three factors, presentation modality, group and use, were significantly different from each other ($p<.0001$). The interaction of group and presentation format was marginally significant ($p<.052$).

Insert Figure 2 about here.

Word Recognition Tests. Figure 3 shows the results of the LNT and MLNT open-set word recognition tests as a function of implant use for both "Stars" (top panels) and "Controls" (bottom panels). The LNT and MLNT each contain two types of test items, "easy" words and "hard" words, which differ in lexical difficulty (Kirk et al., 1995). Differences in performance on the two types of items within each test are used to assess how well a listener can make fine phonetic discriminations among acoustically similar words. Differences in performance between the LNT and MLNT provide a measure of the degree to which the listener is able to use word length to recognize and access words from memory. The LNT consists of short monosyllabic words, the MLNT consists of longer polysyllabic words. In this figure, correct identification of easy and hard words in each test is plotted separately as a function of length of implant use. Easy words are shown with filled bars; hard words are shown with light bars. The LNT results are shown on the left, the MLNT results on the right.

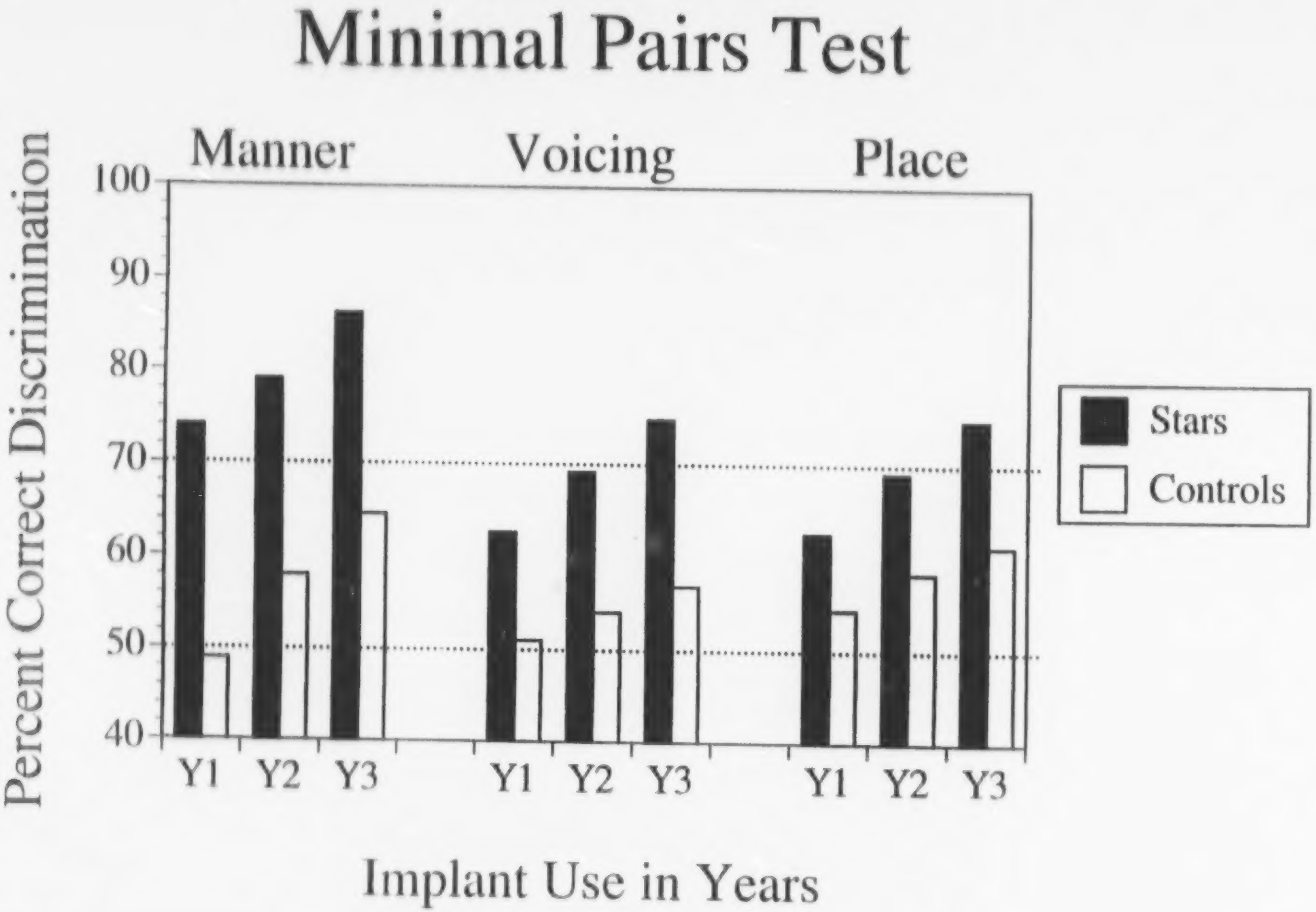


Figure 1. Percent correct discrimination on the minimal pairs (MP) test for manner, voicing and place as a function of implant use. The "Stars" are shown by filled bars, the "Controls" are shown by shaded bars.

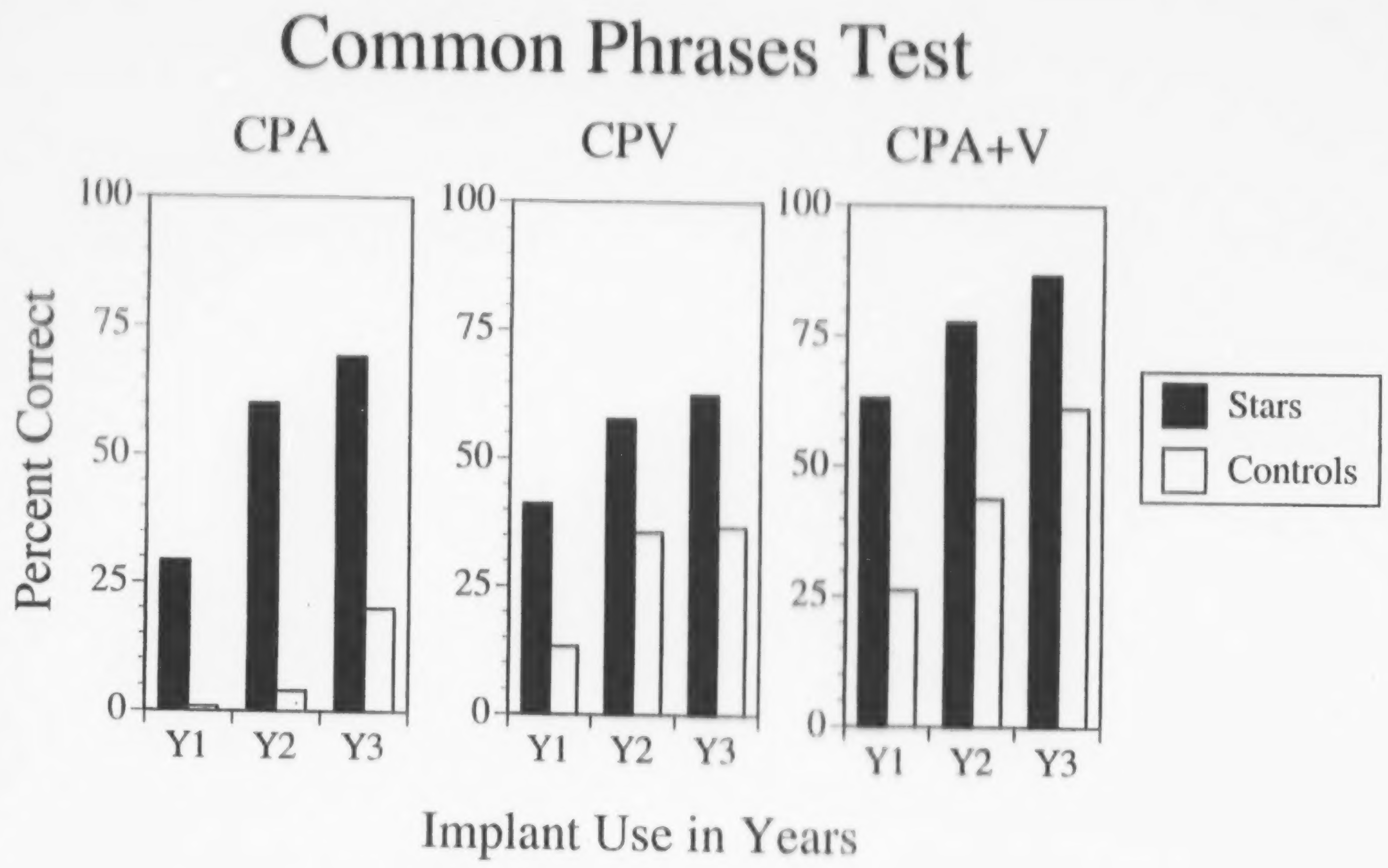


Figure 2. Percent correct performance on the common phrases (CP) test for auditory-only (CPA), visual-only (CPV) and combined auditory plus visual presentation modes (CPA+V) as a function of implant use. The "Stars" are shown by filled bars, the "Controls" are shown by shaded bars.

Insert Figure 3 about here.

Three patterns are shown in this figure. First, the "Stars" consistently demonstrate much higher levels of performance on both the LNT and MLNT words than the "controls." Performance of the "Controls" on both open-set tests was very low and close to the floor even after three years of implant use. Although both groups of subjects were initially selected for this study based on their PBK scores, it is clear from the LNT results that they still maintain these differences with another open-set word recognition test. Thus, the abilities and skills used to recognize words on the PBK test and dissociate the two groups are not specific to the PBK vocabulary items but apparently generalize to another open-set word recognition test, the LNT. The perceptual, cognitive and linguistic skills needed to carry out open-set word recognition tasks may have important diagnostic utility in identifying exceptionally good cochlear implant users. We will return to this issue later in the discussion when we consider the nature of the task demands of these tests.

Second, the results obtained with the "Stars" also show a word-length effect at each testing interval. Performance is better on the MLNT, which contains longer polysyllabic words than on the LNT, which contains only short monosyllabic words. Because the performance levels were so low for the controls, this pattern is obscured by a floor effect for this group. Third, the "Stars" also show an effect of "lexical discrimination" for both the LNT and MLNT vocabularies. That is, the "Stars" recognize easy words better than hard words and this difference gets larger with increases in implant use, especially on the MLNT. Again, because of floor effects, the "Controls" did not display the same pattern of performance across these two conditions.

An analysis of variance was carried out separately on the word recognition scores for the "Stars." The results revealed significant main effects for year ($p<.0001$), word length ($p<.0002$) and communication mode ($p<.0001$). None of the interactions reached significance. Performance on both tests increased over time with implant use. Long words on the MLNT were recognized better than the short words on the LNT. And, oral "Stars" performed better than the TC "Stars." Although there was no main effect for easy vs. hard words, post hoc tests revealed a reliable difference for this variable only with the long words on the MLNT ($p<.04$).

Vocabulary Knowledge. Scores on the Peabody Picture Vocabulary Test (PPVT) are shown in Figure 4. Language quotients are displayed for the "Stars" and "Controls" as a function of implant use. Standardized tests of vocabulary and language development are routinely administered using the child's preferred mode of communication (speech or sign) depending on whether the child is in an Oral-only or TC environment. This procedure was followed here as well.

Insert Figure 4 about here.

The scores for both groups, the "Stars" and "Controls," are very similar overall and inspection of this figure shows little evidence of any change in performance on this test over time. Analysis of variance confirmed these observations. None of the main effects for year or group or any of the interactions with these variables were significant for the PPVT. However, the analysis of variance did reveal a main effect for communication mode on this test. Children from TC environments scored significantly higher on the PPVT than children from oral-only environments ($p<.0001$). This difference did not interact with any of

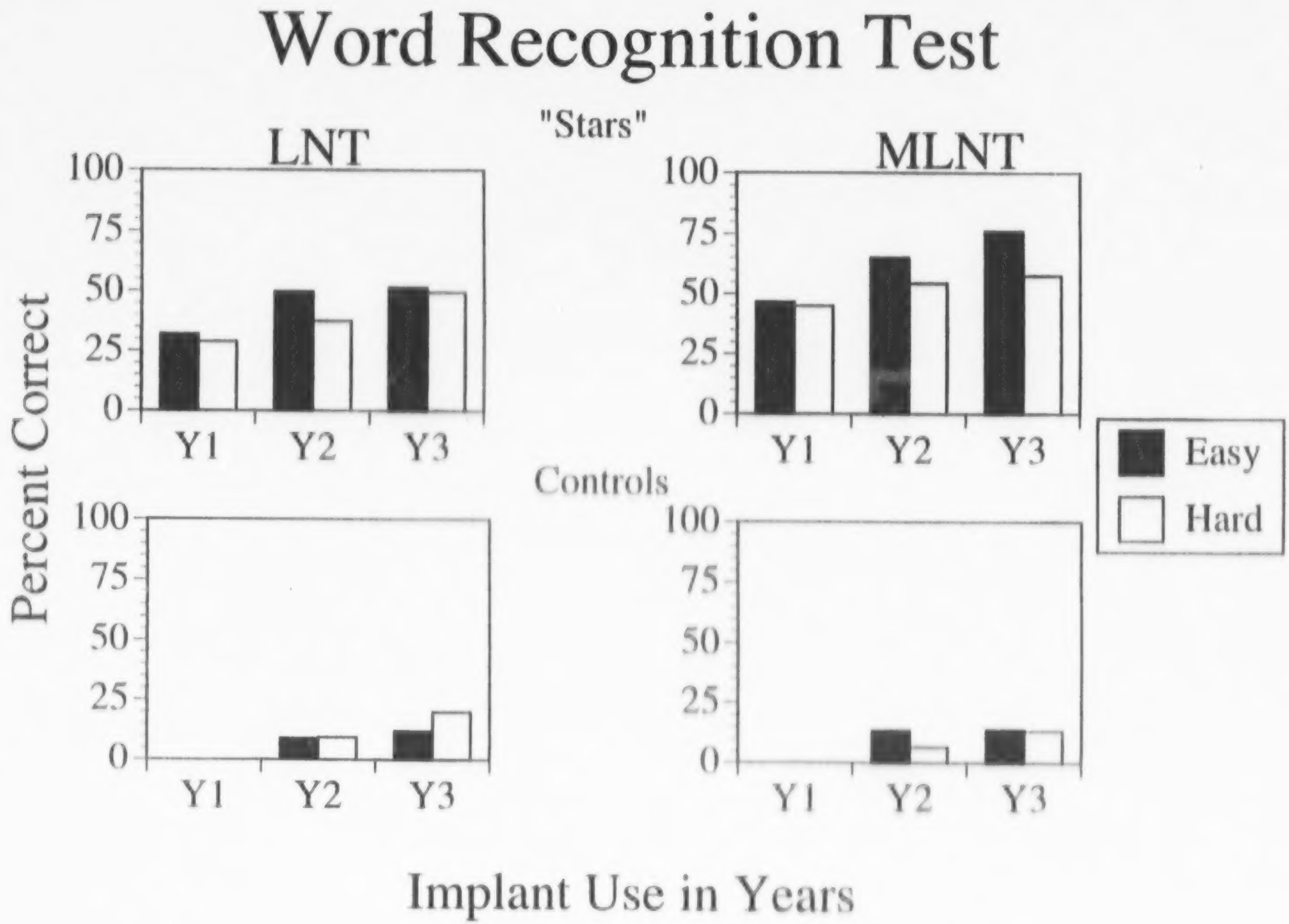


Figure 3. Percent correct word recognition performance for the LNT and MLNT word lists as a function of implant use and lexical difficulty. "Easy Words" are shown by filled bars, "Hard Words" are shown by shaded bars. Data for the "Stars" are displayed in the top panels; the "Controls" are displayed in the bottom panels.

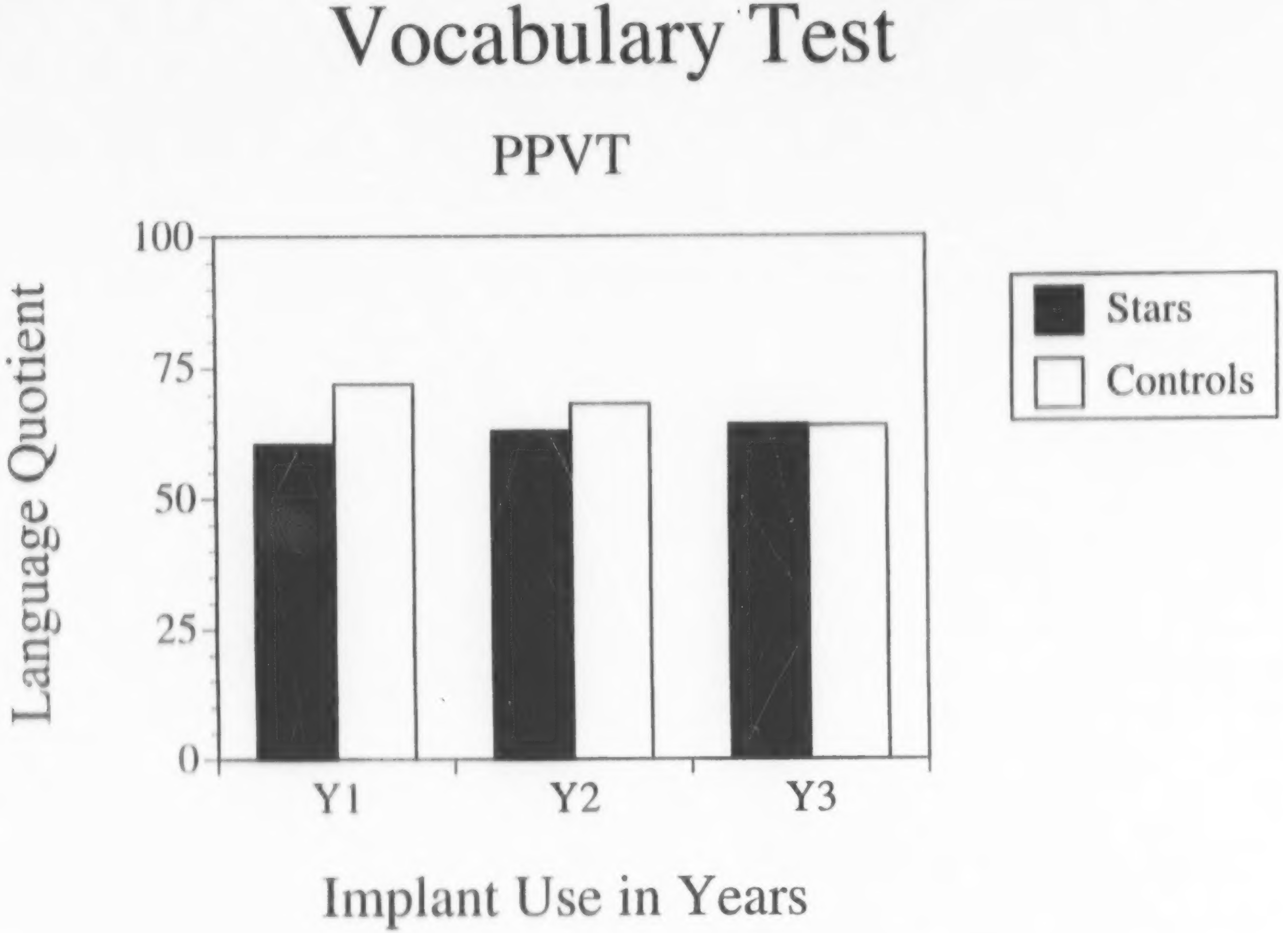


Figure 4. Mean language quotients for the "Stars" and "Controls" on the Peabody Picture Vocabulary Test (PPVT) as a function of implant use.

the other variables and was probably due to the methods used to administer the PPVT to deaf children using their preferred communication mode.

Language Development. Figure 5 shows the Reynell receptive and expressive language scores, expressed as language quotients, for the "Stars" and "Controls" as a function of length of implant use. This figure shows that the differences in performance between these two groups were not very large nor was there any change over time with implant use. As with the PPVT, the Reynell results shown here may also reflect the specific procedures used to administer this test to deaf children. Because the Reynell test is used to assess language development independently of input modality, the child is encouraged to use his/her preferred mode of communication. Thus, for oral children, the test is carried out using speech; for TC children, the test is conducted using sign.

Insert Figure 5 about here.

An analysis of variance on both the receptive and expressive scores of the Reynell revealed two significant main effects and one interaction. The differences between the "Stars" and "Controls" were significant for both the receptive and expressive scales on the Reynell ($p<.0005$). The "Stars" obtained higher language scores than the "Controls" on both tests. However, the analysis also showed a main effect of communication mode ($p<.01$) and an interaction of communication mode with group ($p<.05$). Overall, TC children scored higher than oral-only children. Examination of the interaction between group and communication mode showed that the differences in communication mode were present only for the control subjects, not the "Stars." Although only marginally significant ($p<.09$), the interaction of communication mode with the receptive/expressive variable showed that the TC children were better than oral-only children on the receptive subtest but not on the expressive subtest. Taken together with the PPVT vocabulary scores, these results demonstrate that the preferred communication mode does influence the outcome of standardized tests that assess language and language-dependent abilities such as vocabulary knowledge and language use. The results also suggest that the specific types of linguistic interactions in the child's post-implant environment may play an important role in subsequent language development, vocabulary acquisition and success with a cochlear implant.

Speech Intelligibility. The speech intelligibility scores for the "Stars" and "Controls" are shown in Figure 6 as a function of length of implant use. These data are based on the percentage of correct transcriptions generated for each subject. The speech intelligibility scores are obtained from groups of naïve normal-hearing adult listeners who were asked to listen to and transcribe samples of speech from each subject. The data shown in this figure represent composite scores of three listeners to each child's utterance.

Insert Figure 6 about here.

The results obtained from the transcription task demonstrate that the "Stars" had much better speech intelligibility than the "Controls." Both groups of subjects also showed improvement in speech intelligibility over time. Because the distribution of oral and TC subjects within the "Stars" and "Controls" was unbalanced using the original selection criteria that assigned subjects to these two groups, we also examined the speech intelligibility scores separately for the Oral and TC subjects within each of the groups

Reynell Developmental Language Scales

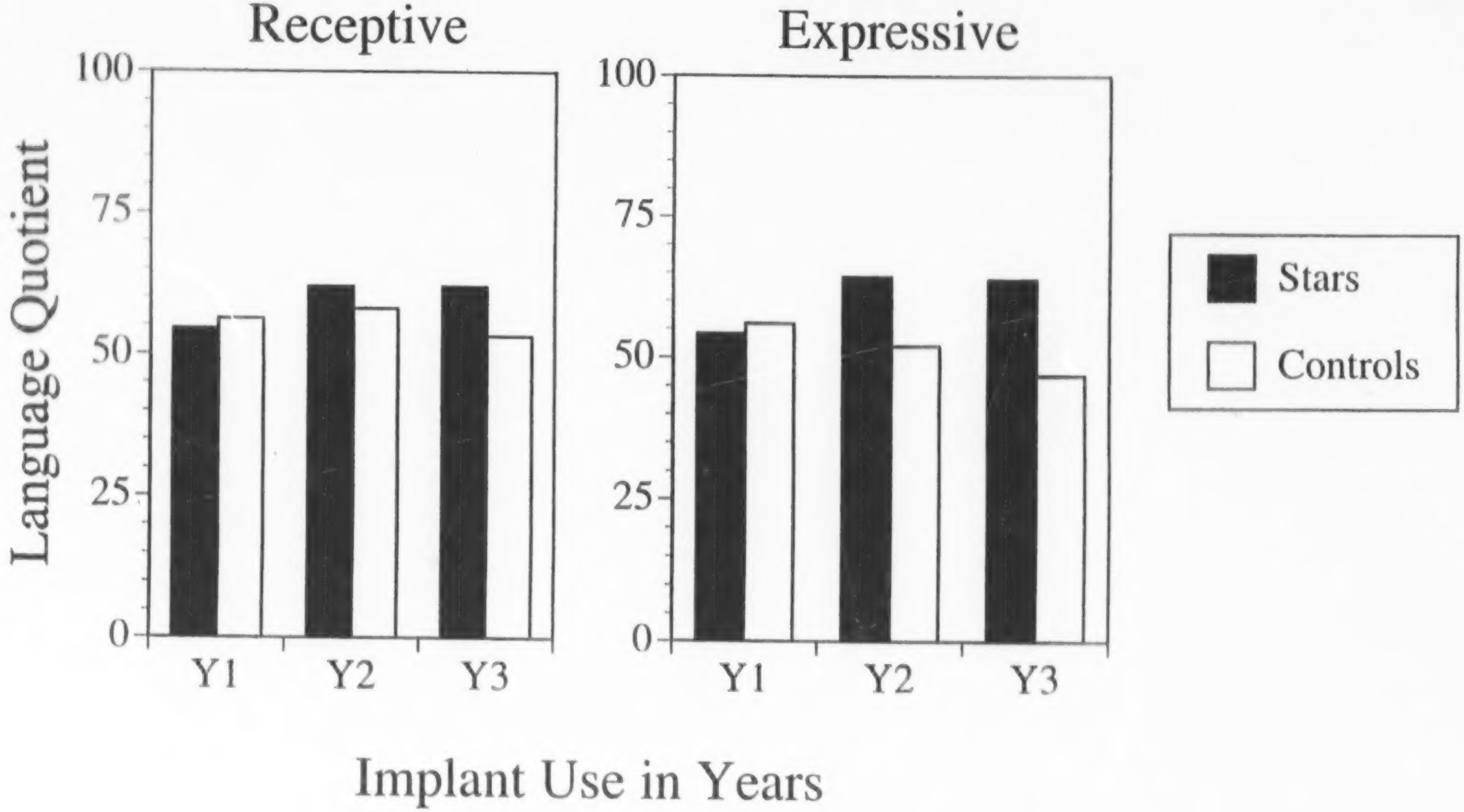


Figure 5. Reynell receptive and expressive language scores for "Stars" and "Controls" as a function of implant use.

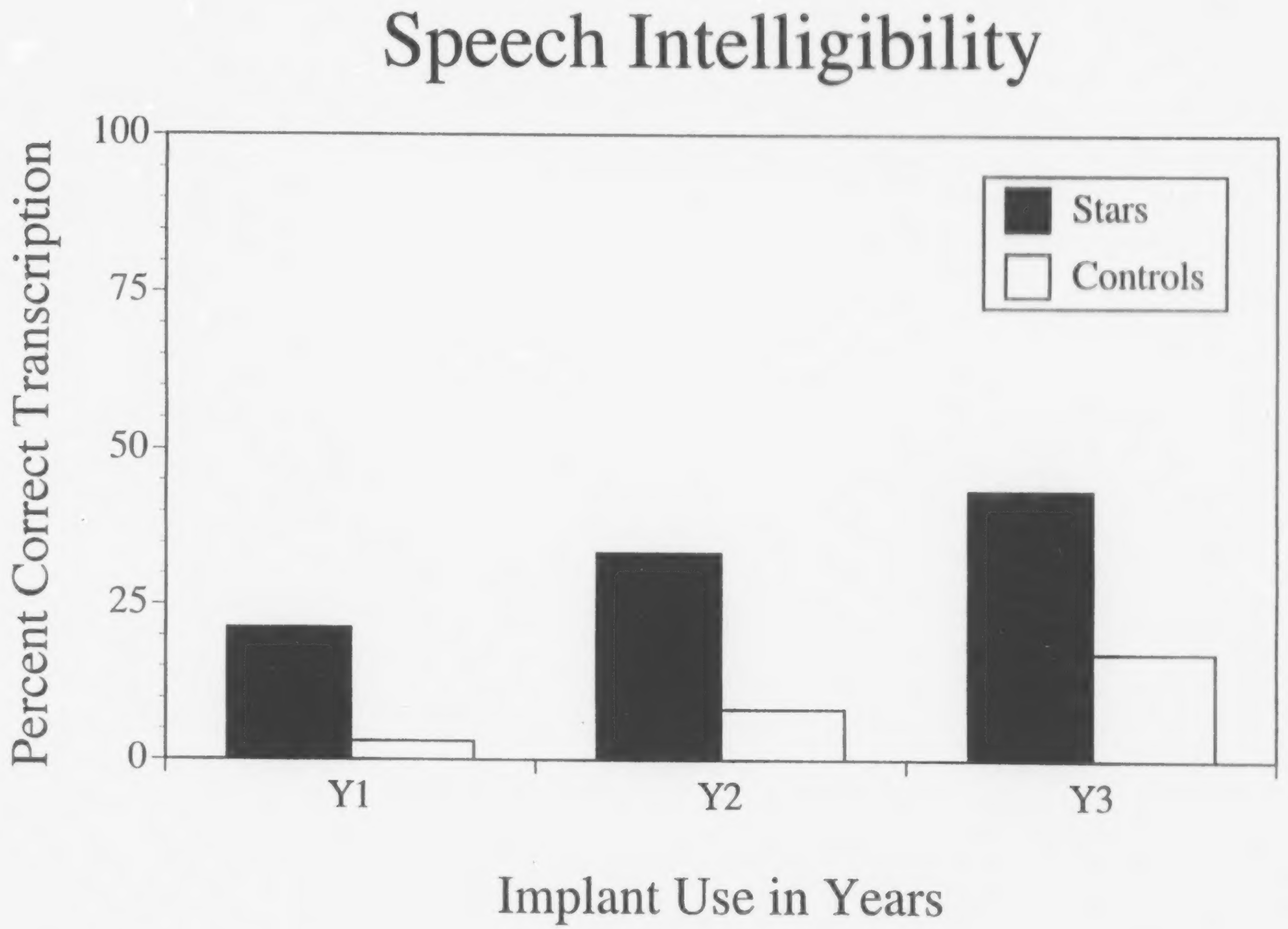


Figure 6. Percent correct transcription scores as a function of implant use for "Stars" and "Controls."

to determine if there were any differences in the speech intelligibility as a function of communication mode. Figure 7 shows a comparison of the "Stars" and "Controls" for oral-only subjects on the left and TC subjects on the right. Although small differences can be seen in overall level of performance between Oral and TC subjects, the same general pattern of performance for the "Stars" and "Controls" is present in both sets of data when the results are examined separately by communication mode.

Insert Figure 7 about here.

An analysis of variance was carried out to assess differences in these variables. The results of this analysis showed only two significant main effects, year ($p<.0001$) and group ($p<.0001$). Speech intelligibility improved with increases in implant use and the "Stars" were consistently better than the "Controls" at each time interval. The main effect of communication mode and all remaining interactions were not significant.

Taken together, the results obtained on tests of speech perception, word recognition, vocabulary, language development and speech intelligibility reveal an interesting and informative pattern of both differences and similarities in performance between the "Stars" and the "Controls." One very important finding observed here was that the "Stars" were not consistently better than the "Controls" across the board on every test measure obtained. Instead, we observed a pattern of "selective" differences between these groups that depended on the particular test and the specific task demands. In addition, we also found effects of communication mode that entered into several interactions with group. Effects of communication mode were observed for the vocabulary and language measures that relied heavily on the child's preferred mode of communication.

For the tests that required the use of speech and spoken language, we found large and consistent differences in performance between the "Stars" and "Controls," and these differences were present regardless of the child's communication mode. Moreover, for these speech-related measures, both the "Stars" and "Controls" showed improvements over time with implant use. Both sets of findings suggest that the differences in performance observed between the "Stars" and "Controls" are not due simply to an overall difference in performance levels or some global predisposition related to intelligence, attention, or cognitive style (Quittner et al., 1994). The differences between these two groups of subjects reflect fundamental differences in how sensory information, specifically information about speech and spoken language, is perceived, encoded in memory, retrieved and subsequently used to perform sequences of operations that require the child to actively manipulate phonological representations of these input signals.

In this connection, it is also important to point out here that although the "Stars" and "Controls" were originally selected based on their performance on the PBK test, the differences in performance between these two groups continue to be observed and maintained on several other speech perception and word recognition tests as well, specifically, the LNT, which is also an open-set test of word recognition. Thus, there is nothing special or unique about the particular words used on the PBK test or the fact that this test is considered to be very difficult for children with cochlear implants (see Meyer & Pisoni this volume). However, we believe there may be something important and informative about the specific task demands of open-set word recognition tests that make use of speech- and language-related skills and abilities that are used in learning spoken language via the auditory modality. In order to identify the underlying factor or set of factors that characterize the exceptionally high levels of performance found with

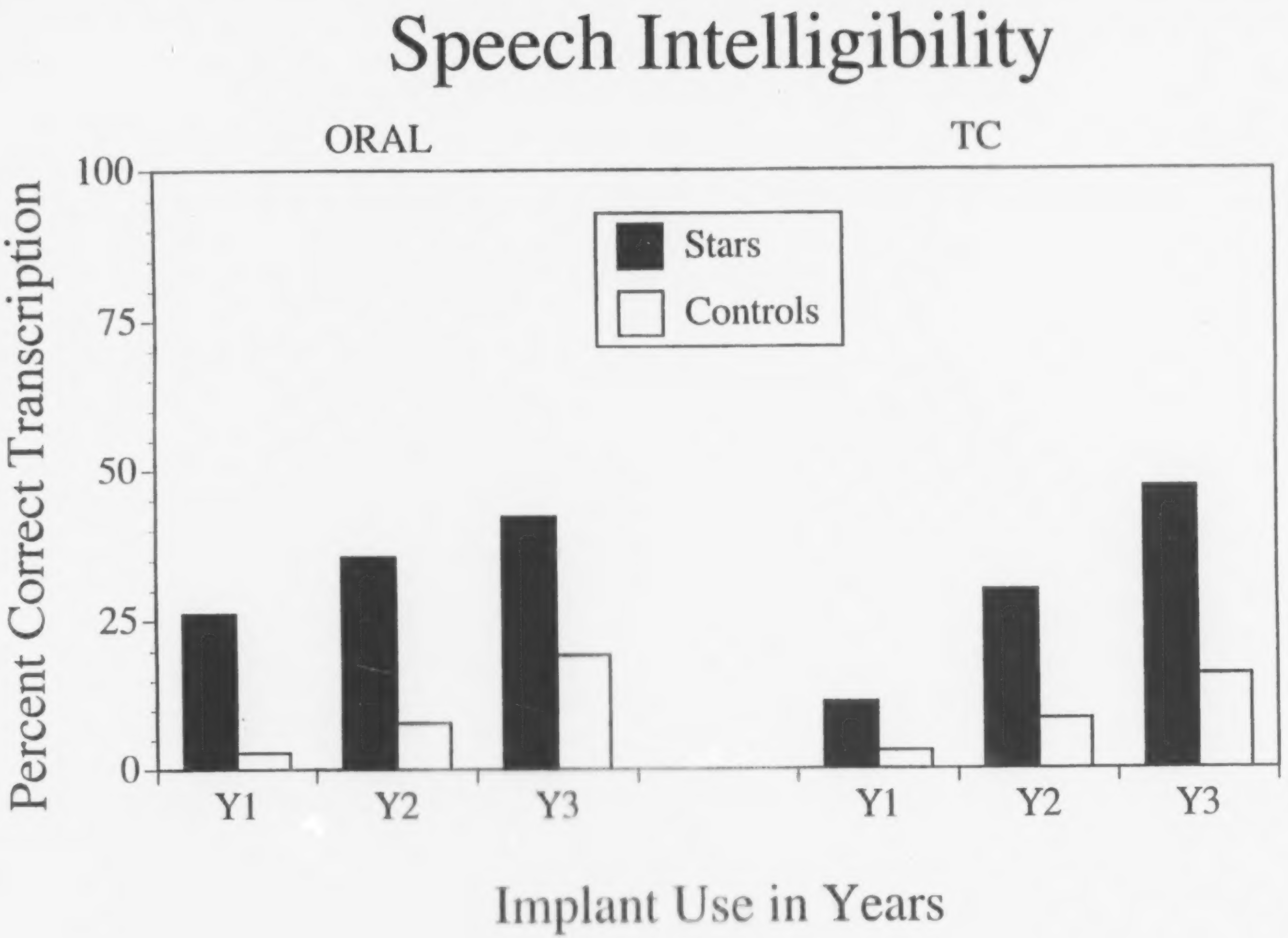


Figure 7. Percent correct transcription scores as a function of implant use for "Stars" and "Controls." Data for oral-only children are shown on the left and total communication (TC) children are shown on the right.

BEST COPY AVAILABLE

the "Stars," we carried out a series of correlations among the various tests. The results of these analyses are reported below in the next section.

Functional Assessments. In addition to the performance data reported in the sections above, we also obtained information on the teacher/parent assessment of the child's meaningful use of speech using two rating scales, the MAIS and the MUSS. The MAIS is used to assess receptive changes in speech perception and spoken language skills; the MUSS is used to assess expressive changes in speech production and language use. Both scales require the respondents to provide a rating for a set of target behaviors. Figure 8 shows the mean ratings for the MAIS in the left panel and the MUSS in the right panel for both the "Stars" and the "Controls" as a function of implant use.

Insert Figure 8 about here.

This figure reveals that both measures increased over time for both the "Stars" and "Controls." The parents' ratings of the "Stars" were, however, consistently higher than the "Controls" at each testing interval for both the MAIS and the MUSS. Analysis of variance confirmed these observations. The increases in the ratings over time with implant use were reliably different for both measures ($p<.0001$), and the difference in the ratings between the "Stars" and "Controls" were significant at each testing interval ($p<.0001$). In addition, this analysis revealed two other main effects, communication mode ($p<.0001$) and test ($p<.007$) as well as a marginal interaction of communication mode and test ($p<.06$). Averaged over the other variables, oral-only children received higher ratings than the TC children, and scores on the MAIS were higher than scores on the MUSS. Examination of the interaction between communication mode and test showed that the ratings for the MAIS were higher than the ratings for the MUSS, but this was found only for the TC children, which may reflect their consistently poorer speech intelligibility and expressive language skills.

Correlations Among Test Measures

To assess the interrelationships among the various test scores for the measures obtained in this study, a series of simple correlations were carried out separately for the "Stars" and "Controls" for each of the dependent measures described earlier. All the correlations reported below were computed on test scores obtained after only one year of implant use in order to determine if we can identify the exceptionally good users at an early point in time.

Minimal Pairs Test. Table II shows the correlations for the three speech features—manner, voicing and place—on the Minimal Pairs Test and the other test scores. The table reveals several consistent patterns among the correlations. First, looking at the overall results for the "Stars," it is clear that a large number of the correlations with the Minimal Pairs Test are positive and quite strong. This pattern is particularly impressive given the small sample size available for this study.

87

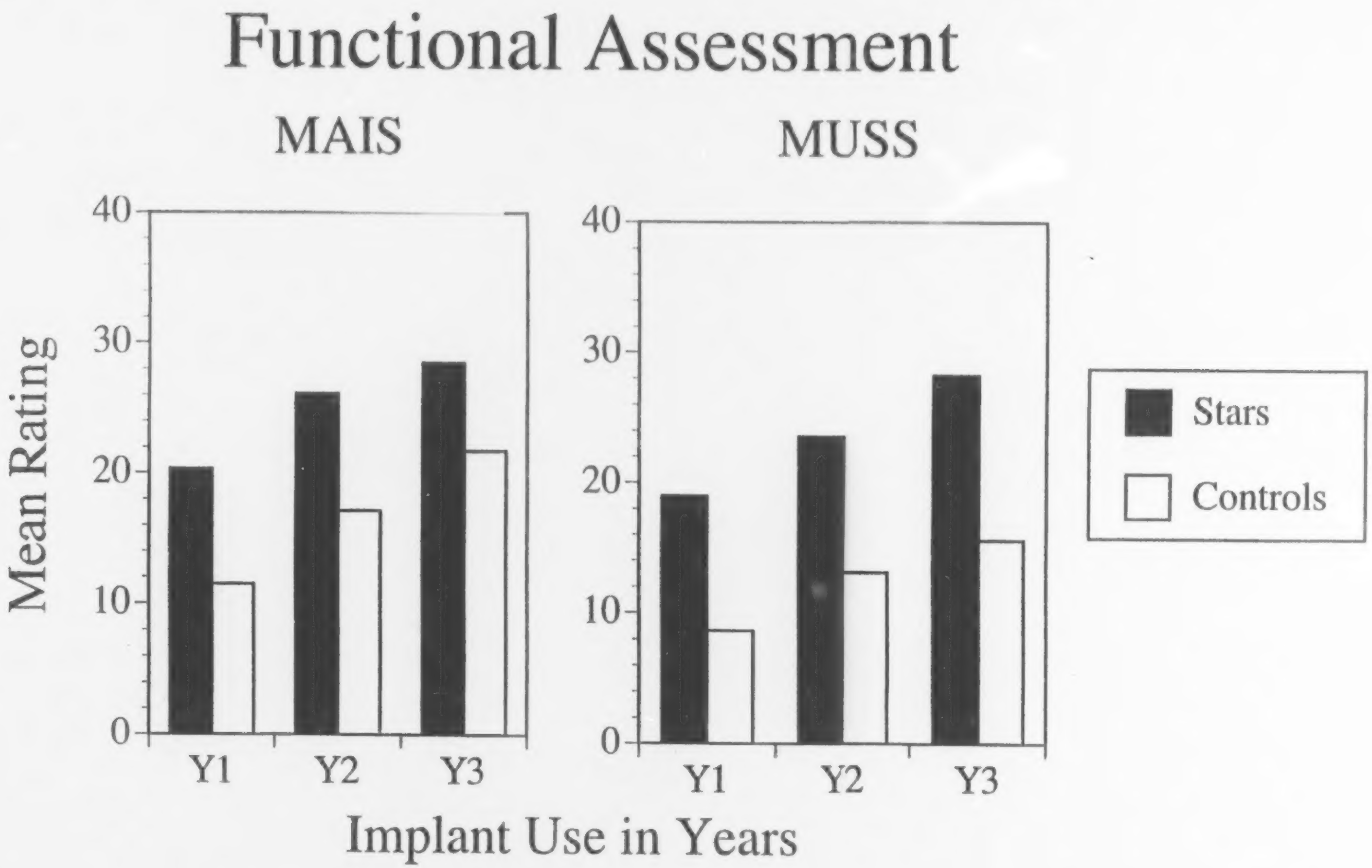


Figure 8. Mean functional assessment ratings by parents for the MAIS, a receptive language scale, shown on the left and the MUSS, an expressive language scale, shown on the right as a function of implant use. The "Stars" are shown by the filled bars; the "Controls" by the shaded bars.

TABLE II

CORRELATIONS: SPEECH PERCEPTION - YEAR 1

	Minimal Pairs Test					
	Manner		*Voicing*		*Place*	
	Stars	Controls	Stars	Controls	Stars	Controls
COMPREHENSION:	r =	r =	r =	r =	r =	r =
Common Phrases-Auditory only	.58**	-.51	.38	-.05	.04	.46
Common Phrases-Visual only	.80***	-.19	.51*	-.70	.18	-.19
CommonPhrases-Auditory+Visual	.86***	-.38	.53*	-.46	.16	-.03
WORD RECOGNITION:						
LNT-Easy Words	.34	—	.20	—	.16	—
LNT-Hard Words	.51	—	.58	—	-.06	—
MLNT-Easy Words	.53	—	.34	—	.06	—
MLNT-Hard Words	.38	—	.33	—	-.11	—
VOCABULARY:						
PPVT	.56**	.20	.46*	.23	.07	-.25
LANGUAGE:						
Reynell Receptive Language Quotient	.77**	.08	.69*	-.63	.20	.01
Reynell Expressive Language Quotient	.78**	-.28	.61*	-.49	.31	.33
SPEECH INTELLIGIBILITY:						
Transcription	.55	.19	.53	-.11	.41	-.09

* $p < .05$
** $p < .01$
*** $p < .001$

Manner discrimination was correlated significantly with language comprehension performance as indexed by all three formats of the Common Phrases Test. Vocabulary knowledge, the receptive and expressive subtests of the Reynell Language Scales and speech intelligibility were also correlated significantly with manner discrimination on this test.

Discrimination of the voicing feature for the "Stars" on the Minimal Pairs Test was also positively correlated with several of the other test scores. The overall pattern for voicing was very similar to the one observed for manner, although the correlations for voicing were slightly lower.

In contrast to the results observed for manner and voicing, the correlations obtained for the place feature on the Minimal Pairs Test shown in this table were generally quite low and none of them were statistically significant. The absence of correlations of place with the other test scores reflects the very poor level of discrimination of place contrasts on the Minimal Pairs Test. With performance close to chance on the place feature, as shown earlier in Figure 1, there was little room for the scores to vary and therefore the observed correlations were very low.

A very different pattern of correlations with these test scores can be seen for the control group in Table II. Overall, these correlations were much lower and almost all of them were non-significant. Finally, there were several cases where the sample size was simply too small to report correlations. These are shown by dashes in the table and represent the absence of scores on the LNT and MLNT word recognition tests.

Taken together, the pattern of correlations obtained with the Minimal Pairs Test for manner and voicing features suggests that the "Stars" are not only able to discriminate small differences between minimal pairs of spoken words but they are also able to encode these differences and make use of the available sensory input in other language-related tasks. The control subjects are unable to do this above chance levels of performance. The ability to discriminate spoken words even based on broad phonetic categories using only manner and voicing features and encode them in memory in the form of phonological representations may be sufficient to support a variety of receptive and expressive language functions in the absence of additional, possibly redundant sensory information.

The low correlations found among these test measures in the control group suggest that the discrimination of small differences among minimal pairs of spoken words may play an extremely important role in predicting more complex language-dependent abilities as the sensory input is encoded and propagates up the information processing system. Although the "Stars" are apparently unable to discriminate and reliably use place information until three years post-implant, they appear able to make efficient use of the manner and voicing differences among spoken words to discriminate reliably in the two-alternative forced-choice minimal pairs test. Partial acoustic-phonetic information about the sound patterns of words may be sufficient at this point in development to support word recognition and lexical access and to permit the retrieval and implementation of sensory-motor plans used in speech production. It is interesting to note that the correlations of manner and voicing with word recognition scores on the LNT for the "Stars" were also positive and strong, although they did not reach statistical significance.

Common Phrases Test. A summary of the correlations for the Common Phrases Test and the other test scores is shown in Table III. The results are displayed separately for the three presentation formats, auditory-only (CPA), visual-only (CPV), and auditory+visual (CPAV). Examination of the correlations for the "Stars" also reveals a pattern of very strong positive correlations of the Common Phrases Test with the other test measures, particularly for in the auditory-only conditions. Performance on the auditory-only condition for the "Stars" was significantly correlated with almost all of the dependent measures shown in this table. The only exceptions were the voicing and place features of the Minimal Pairs Test. This pattern of results strongly suggests a common underlying source of variance that is shared among all these tasks. Most noteworthy were the extremely high correlations of CPA with the LNT word recognition test and both the receptive and expressive measures of language development assessed by the Reynell Developmental

Language Scales. These correlations are extremely high and statistically significant even given the modest sample size used here.

TABLE III

CORRELATIONS: COMPREHENSION - YEAR 1

	Common Phrases Test					
	Auditory-only		*Visual-only*		*Auditory+Visual*	
	Stars	Controls	Stars	Controls	Stars	Controls
SPEECH PERCEPTION:	r =	r =	r =	r =	r =	r =
Minimal Pairs-Manner	.58**	-.51*	.80***	-.19	.86***	-.38
Minimal Pairs-Voicing	.31	-.05	.51*	-.70	.53*	-.46
Minimal Pairs-Place	.04	.46	.18	-.19	.16	-.03
WORD RECOGNITION:						
LNT-Easy words	.81***	----	.41	----	.42	----
LNT-Hard words	.85***	----	.57	----	.56	----
MLNT-Easy words	.83***	----	.60*	----	.56*	----
MLNT-Hard words	.70*	----	.62*	----	.33	----
VOCABULARY:						
PPVT	.69***	-.46	.40	-.47	.50*	.04
LANGUAGE:						
Reynell Receptive Language Quotient	.82***	----	.64*	----	.64*	.33
Reynell Expressive Language Quotient	.85***	----	.79**	----	.67*	.36
SPEECH INTELLIGIBILITY:						
Transcription	.65*	.04	.87***	.25	.43	.07

* $p < .05$
** $p < .01$
*** $p < .001$

The Common Phrases Test is considered to be a difficult open-set test of spoken language comprehension. The child is required to name objects and carry out a series of commands or instructions. The test is administered in three presentation formats. Although the correlations of each of the three formats with the other test measures are positive and significant for the "Stars," it should be noted here that the strongest correlations were observed in the auditory-only presentation condition (CPA). The only exception to this observation were the correlations of common phrases visual-only (CPV) with speech intelligibility. Here we see a very high correlation (r= +.87, which suggests that the "Stars" may also be encoding and using information about the visual consequences of speech production to guide their own speech articulation. Thus, the "Stars" are not only sensitive to the auditory properties of speech encoded in the speech waveform, but they are also able to encode and use phonetically-relevant information about the talker's articulation present in the optical display of the talker's face and lips. The use of multi-modal information about the phonetically distinctive properties of speech may be another diagnostic feature that characterizes the "Stars" and distinguishes them from other cochlear implant users. The speech reading skills of the "Stars" may be superior to the "Controls" not only because they are able to use the additional visual information to recognize words already encoded in memory but also because they can integrate different forms of a common phonetic event that are generated from the same source, namely, the speaker's vocal tract.

Word Recognition Tests. Tables IV-A and IV-B display the correlations for the LNT and MLNT word recognition tests with each of the other test measures. The correlations for the easy words are shown in the left columns and the hard words are shown in the right columns of each table for the "Stars." Because the scores for the control subjects were so low on these open-set tests, it was not possible to compute correlations for this group with the other test measures.

TABLE IV-A

CORRELATIONS: WORD RECOGNITION - YEAR 1

	Lexical Neighborhood Test (LNT)			
	Easy Words		*Hard Words*	
	Stars	Controls	Stars	Controls
SPEECH PERCEPTION:	r =	r =	r =	r =
Minimal Pairs-Manner	.34	----	.51	----
Minimal Pairs-Voicing	.20	----	.58	----
Minimal Pairs-Place	.16	----	-.06	----
COMPREHENSION:				
Common Phrases-Auditory only	.81***	----	.85***	----
Common Phrases-Visual-only	.41	----	.57	----
Common Phrases-Aditory+Visual	.42	----	.55	----
VOCABULARY:				
PPVT	.62*	----	.63*	----
LANGUAGE:				
Reynell Receptive Lanaguage Quotient	.86***	----	.81**	----
Reynell Expressive Lanaguage Quotient	.83***	----	.82**	----
SPEECH INTELLIGIBILITY:				
Transcription	.89**	----	.80**	----

* $p < .05$
** $p < .01$
*** $p < .001$

TABLE IV-B

CORRELATIONS: WORD RECOGNITION - YEAR 1

Multisyllabic
Lexical Neighborhood Test
(MLNT)

	Easy Words		*Hard Words*	
	Stars	Controls	Stars	Controls
SPEECH PERCEPTION:	r =	r =	r =	r =
Minimal Pairs-Manner	.34	—	.33	—
Minimal Pairs-Voicing	.53	—	.38	—
Minimal Pairs-Place	.06	—	-.11	—
COMPREHENSION:				
Common Phrases-Auditory	.83**	—	.70*	—
Common Phrases-Visual-only	.60*	—	.62*	—
Common Phrases-Auditory+Visual	.56*	—	.33	—
VOCABULARY:				
PPVT	.57*	—	.35	—
LANGUAGE:				
Reynell Receptive Language Quotient	.84**	—	.66*	—
Reynell Expressive Language Quotient	.87***	—	.76**	—
SPEECH INTELLIGIBILITY:				
Transcription	.87**	—	.72	—

* $p < .05$
** $p < .01$
*** $p < .001$

Examination of the correlations in both tables shows a very similar pattern of results. Performance on both word recognition tests is highly correlated with comprehension scores, vocabulary knowledge, language development as well as speech intelligibility. The correlations are very strong and statistically significant. The correlations of the LNT and MLNT with performance on the Minimal Pairs Test were much smaller and none of them reached significance.

The pattern of correlations for the "Stars" shown here for both word recognition tests is very similar to the overall pattern of correlations observed earlier with the other test measures, again suggesting a common underlying source of variance in these tasks. The extremely high correlations of the word recognition scores with Common Phrases-Auditory Only and both language measures on the Reynell suggest that this common source of variance may be related in some way to the encoding, storing and retrieving of spoken words and access to phonological information about spoken words in lexical memory. Whatever the precise description of this source of variance turns out to be, it is clear from the correlations shown in these two tables that it is related to the perception of words in an open-set testing format. The processes used to identify words in an open-set test like the PBK or the LNT are also used in other language-related tasks such as comprehension and speech production and it is these particular cognitive and linguistic abilities and skills that seem to be well-developed and easily deployed/accessed by the "Stars" in these tasks.

Language Development. The correlations between the receptive and expressive scales on the Reynell Developmental Language Scales and the other dependent measures are shown in Table V for the "Stars" and "Controls." This table shows once again a very systematic pattern of correlations among the test scores.

TABLE V

CORRELATIONS: LANGUAGE - YEAR 1

Reynell Language Scales
(Language Quotient)

	Receptive		*Expressive*	
	Stars	Controls	Stars	Controls
SPEECH PERCEPTION:	r =	r =	r =	r =
Minimal Pairs-Manner	.77**	.08	.78**	-.28
Minimal Pairs-Voicing	.69*	-.63	.61*	-.49
Minimal Pairs-Place	.20	-.01	.31	.33
COMPREHENSION:				
Common Phrases-Auditory	.82**	—	.85***	—
Common Phrases-Visual-only	.64*	—	.79**	—
Common Phrases-Auditory+Visual	.64*	.33	.67*	.36
WORD RECOGNITION:				
LNT-Easy words	.86***	—	.83***	—
LNT-Hard words	81**	—	.82**	—
MLNT-Easy words	.84**	—	.87***	—
MLNT-Hard words	.66*	—	.76	—
VOCABULARY:				
PPVT	.81***	.69**	.68**	.56*
SPEECH INTELLIGIBILITY:				
Transcription	.80**	-.39	.85**	-.13

* $p < .05$
** $p < .01$
*** $p < .001$

Performance on both the receptive and expressive scales of the Reynell is highly correlated with each of the other dependent measures. The correlations are extremely high and they are all statistically significant. The very strong correlations of the Reynell scores with the word recognition tests provide additional support for the hypothesis that the common underlying factor that is operating here is related in some way to the encoding, storage, retrieval and manipulation of spoken words. Our proposal concerning the importance and central role of spoken word recognition and lexical access also receives additional support from the presence of very strong correlations with speech intelligibility shown here and in the previous tables for the LNT and MLNT. Correlations between measures of speech perception and speech production reflect the transfer of linguistic knowledge from one modality to another and suggest a common locus or shared representational system that is used for both receptive and expressive language functions (Prince & Smolensky, 1997). This shared representational system has the properties of a grammar of the language that the child is exposed to in the ambient linguistic environment.

Speech Intelligibility. The correlations of the speech intelligibility scores and other measures are shown in Table VI for the "Stars" and "Controls." Examination of the correlations among these variables shows the same consistent pattern that was observed in the previous tables for the perceptual tests.

TABLE VI

CORRELATIONS: SPEECH INTELLIGIBILITY - YEAR 1

	Transcription Scores	
	Stars	Controls
SPEECH PERCEPTION:	r =	r =
Minimal Pairs-Manner	.55	.19
Minimal Pairs-Voicing	.53	-.11
Minimal Pairs-Place	.41	-.09
COMPREHENSION:		
Common Phrases-Auditory	.65**	.04
Common Phrases-Visual-only	.87**	.25
Common Phrases-Auditory+Visual	.43	.07
WORD RECOGNITION:		
LNT-Easy Words	.89**	—
LNT-Hard Words	.80*	—
MLNT-Easy Words	.87**	—
MLNT-Hard Words	.72	—
VOCABULARY:		
PPVT	.45	-.01
LANGUAGE:		
Reynell Receptive Language Quotient	.80**	-.39
Reynell Expressive Language Quotient	.85**	-.13

* $p < .05$
** $p < .01$

BEST COPY AVAILABLE

Speech intelligibility scores for the "Stars" are also highly correlated with spoken language comprehension, word recognition and language measures. The correlations of speech intelligibility with the Minimal Pairs Test and vocabulary knowledge as measured by the PPVT were also positive, although they were not as strong as correlations with the other measures and did not reach statistical significance.

Functional Assessments. Table VII provides a summary of the correlations for the two functional assessments, the MAIS and the MUSS, with the other test measures.

TABLE VII

CORRELATIONS: FUNCTIONAL ASSESSMENTS - YEAR 1

	MAIS [a]		*MUSS* [b]	
	Stars	Controls	Stars	Controls
SPEECH PERCEPTION:	r =	r =	r =	r =
Minimal Pairs-Manner	.19	.44	.47	-.31
Minimal Pairs-Voicing	.32	.03	.27	-.50
Minimal Pairs-Place	-.25	.37	-.27	.62
COMPREHENSION:				
Common Phrases-Auditory	-.07	.15	.26	----
Common Phrases-Visual-only	-.05	----	.60*	----
Common Phrases-Auditory+Visual	.06	----	.40	----
WORD RECOGNITION:				
LNT-Easy Words	.07	----	.31	----
LNT-Hard Words	.19	----	.31	----
MLNT-Easy Words	.21	----	.57	----
MLNT-Hard Words	.13	----	.53	----
VOCABULARY:				
PPVT	.20	-.35	.15	-.21
LANGUAGE:				
Reynell Receptive Language Quotient	.17	-.31	.34	-.14
Reynell Expressive Language Quotient	.27	-.21	.61*	.29
SPEECH INTELLIGIBILITY:				
Transcription	-.05	.61*	.18	.54

* $p < .05$
** $p < .01$
*** $p < .001$

[a] Meaningful Auditory Integration Scale (Robbins & Osberger, 1991)
[b] Meaningful Use of Speech Scale

BEST COPY AVAILABLE

Examination of the correlations for both the "Stars" and "Controls" shows a pattern that is quite different from the previous tables of correlations. Except for a few random cases, almost all of the correlations with the functional assessments shown in this table are quite low and not statistically significant. This pattern contrasts markedly with the previous results. The very high intercorrelations of measures of speech perception, spoken language comprehension, word recognition and language observed earlier for the "Stars" was not replicated here either for the MAIS, which was designed to measure parents' assessment of their child's receptive skills, or the MUSS, which measures the parents' judgments of the child's expressive skills in using spoken language.

The absence of strong positive correlations with the functional assessment measures shown here is interesting because reliable differences in these ratings were found between the "Stars" and "Controls." Moreover, the ratings improved over time with implant use. The failure to find correlations with these measures may reflect the fact that the two sources of variance come from entirely different distributions. The speech and language performance measures were obtained directly from the "Stars" and "Controls" whereas the ratings on the MAIS and MUSS were based on the parents' subjective reports of their child's receptive and expressive behaviors.

General Discussion

Until the present investigation, there has been little previous research directed specifically at the study of individual differences among pediatric cochlear implant users or an examination of the perceptual, cognitive and linguistic abilities of the exceptionally good subjects, the "Stars." The results of our analyses of measures of speech perception, word recognition, spoken language comprehension, vocabulary knowledge and language development demonstrate that a child who displays exceptionally good performance on the PBK test also shows very good scores on a variety of other speech and language measures as well. This is a revealing and theoretically important finding. The differences in performance observed here between the "Stars" and "Controls" are of substantial theoretical interest because it may now be possible to determine precisely how and why the "Stars" differ from other less successful cochlear implant users. If we have knowledge of the factors that are responsible for individual differences in performance among deaf children who receive cochlear implants, particularly the variables that underlie the extraordinarily good performance of the "Stars," we may be able to help those children who are not doing as well with their implant at an early point in development. Moreover, these findings may have direct clinical relevance in terms of recommending specific changes in the child's language-learning environment and in modifying the nature of the sensory inputs and linguistic interactions a child has with his/her parents, teachers and speech therapists who provide the primary language model for the child. Our findings on individual differences may also help in providing clinicians and parents with a principled basis for generating realistic expectations about outcome measures, particularly measures of speech perception, comprehension, language development and speech intelligibility in deaf children with cochlear implants.

Performance on Other Tests

The present results also demonstrate that the "Stars" do exceptionally well on another open-set test of spoken word recognition, the LNT. This particular finding suggests that the exceptionally good performance of the "Stars" on the PBK test, which was originally used as the criterial variable to select subjects for inclusion in this study, generalizes to another open-set test of word recognition and to several other behavioral measures that also make use of the same underlying cognitive and linguistic processes. The exceptionally good performance of the "Stars" is therefore not an isolated or anomalous finding that is specific to the particular words used on the PBK test but instead may have something to do more generally with the specific task demands and information processing requirements of open-set tests of speech perception and spoken language comprehension. The pattern of intercorrelations found here suggests that the specific cognitive processes needed to perform these tasks are also recruited and used in other speech and language tasks as well, particularly tasks that require the subject to recognize spoken words and access phonological and lexical representations of words from memory.

Our findings also demonstrate that the "Stars" are not always better than the "Control" subjects on every one of the measures we examined. Performance on tests of vocabulary knowledge, receptive language, non-verbal intelligence, visual-motor integration and visual attention were not significantly different for the two groups compared in this study. Taken together with the pattern of correlations observed among the test measures, the overall results suggest that the differences between the "Stars" and "Controls" may not be due to some global factor related to sensory or cognitive abilities (Smith & Katz, 1996) or to a predisposition to use and exploit partial stimulus information to reconstruct complex patterns from fragments (Watson, 1991). Instead, the differences observed in this study appear to reflect much more selective differences in the processing of auditory information through the cochlear implant. These differences in information processing (i.e., perceiving, encoding, storing and retrieving) are domain-specific in nature and emerge as a function of interactions with the sensory, perceptual and cognitive demands of the language-learning environment to which these children are exposed during the first year after receiving an implant (Newport, 1990).

Ideally, it would be desirable to have a set of pre-implant measures for each child that could be used to predict the outcome measures of speech perception, spoken word recognition, and language development after implantation (Ruben, 1992). However, in acquiring language, it is quite possible and perhaps even very likely that the observed differences among children in speech perception and language-related measures are due to a set of perceptual and cognitive factors that become operative only after implantation takes place (Jusczyk, 1997; in press). This is true because of the important role of early sensory and perceptual experience and the presence of a critical period in development of language (Snow & Hoefnagel-Hoehle, 1978). The failure to find a relationship between pre-implant and outcome measures points to the important role of interactions between processing activities and stimulation that occur in the language-learning environment and the child's newly acquired sensory predispositions to perceive and process sound and speech in meaningful ways (Jusczyk & Aslin, 1995; Saffran, Aslin, & Newport, 1996). The present results demonstrate that these kinds of interactions take place very early on during the first year and continue to at least three years post-implantation.

Sources of Variance

A substantial amount of the variance among subjects may be due to factors such as perceptual learning, attention, memory, and coding, and an interaction between perceptual and cognitive activities in the ambient language learning environment and the sensory capabilities of these children (Seidenberg, 1997). Because researchers have been unable to identify pre-implant measures that can reliably predict post-implant performance on a variety of tests, or account for the substantial individual differences among users, it may be necessary to shift the emphasis of research to the study of psychological variables and cognitive factors such as perception, attention, learning, memory and categorization all of which concern the post-implant environment of child. The previous emphasis on demographic variables such as age at implantation or length of device use should be refocused because these factors are not associated with the underlying sources of variance responsible for the individual differences observed among cochlear implant users. Traditional audiological criteria and methods of hearing assessment are inadequate to study these psychological variables or measure their contribution, because a substantial portion of the variance among subjects comes from information processing operations that are associated with encoding, storage and retrieval of information from memory and from interactions in the language-learning environment.

The large and consistent differences observed between the two groups of subjects on tests that rely on oral language skills requiring recognition of spoken words and the changes in performance with implant use over time on several tests that draw on these same resources, strongly suggests that some form of learning is occurring during the first year after implantation (Pinker, 1991; Jusczyk, 1997). Thus, differences in perceptual learning may be another important factor that predicts success with a cochlear implant. Precisely what is learned during the first year of use after implantation and exactly what kind of learning processes are involved during this time are fundamental questions that we believe are central to understanding the individual differences that have been reported consistently in the pediatric cochlear implant literature over the years. Issues involving learning and categorization have not been investigated in the past, although they may play an extremely important role in understanding how and why some deaf children with cochlear implants are able to show such extraordinarily good performance on several related tasks and why other children fail to reach these speech and language milestones. As the present set of analyses have shown, these differences occur very early after implantation during the first year of use.

The results of the present investigation suggest several hypotheses that can explain the differences in performance between the "Stars" and the "Controls." We believe these accounts are worth pursuing and evaluating in much greater depth, because they suggest new and unexplored areas of basic and clinical research on pediatric cochlear implant users. One explanation places the locus of the differences between the "Stars" and "Controls" at central rather than peripheral processes. This approach is concerned with how sensory information is encoded, stored, retrieved and manipulated in various kinds of information processing tasks such as speech feature discrimination, phoneme identification, word recognition, language comprehension and speech production. The emphasis here is on the perceptual and cognitive factors that play a critical role in how the initial sensory input is processed, organized and used in various psychological tasks. One of the key components that link these various processes and operations together is the working memory system, and it is the properties of this particular memory system that may provide new insights into the nature and locus of the individual differences observed among users of cochlear implants (see Carpenter, Miyake & Just, 1994; Baddeley, Gathercole, & Papagno, 1998). Unfortunately, at this time, we do not have any working memory data for the "Stars" and "Controls" to test this proposal, but several new studies are currently underway with a variety of stimulus materials and experimental methodologies (Cleary, this volume).

One common source of variance that we have been able to identify from the available data appears to be related to the perception, encoding, storage, retrieval and processing of spoken words (Gathercole, 1995; Gathercole, Willis, Baddeley & Emslie, 1994). A very consistent pattern of intercorrelations was observed repeatedly across all the test measures for the "Stars" which was either absent or obscured because of floor effects in the "Control" subjects. This particular source of variance is present and manifests itself most prominently in open-set tests of word recognition and language comprehension such as the PBK, LNT and Common Phrases. Recognizing and retrieving spoken words from memory is also a central component of both subtests of the Reynell Developmental Language Scales and is mandatory in the speech production task as well, where sensory-motor plans for the articulation of words must be retrieved from the lexicon and recruited in speech production.

One major reason the PBK test has been very useful in identifying the exceptionally good users of cochlear implants is that this particular test is an open-set test of word recognition that requires the child to access words from memory based on an analysis of the sensory input in the signal (Gathercole, 1995). In order to access and retrieve words, a child must use a set of processes to initially encode a spoken word in isolation without any surrounding context or response constraints and then retrieve information from his/her lexicon about the properties of this word that can serve as an articulatory plan or a motor program in speech production to produce a verbal response (Beckman & Edwards, in press).

These various perceptual and cognitive operations on spoken words require access to a variety of memory codes and representations of speech and spoken language at different levels of analysis. The speed and efficiency of these information-processing operations, particularly as they might be employed in tasks requiring transformation and mapping from perception to production, will depend to a large extent on having phonetic and phonological representations of words in memory and organizing these representations systematically in a lexicon that can be accessed efficiently to provide different sources of information about the words in the language. The very high correlations observed between the word recognition scores and measures of speech intelligibility suggest that both tasks, although reflecting different aspects of spoken language, are drawing on a common underlying source of variance that reflects the same fundamental information processing operations regardless of modality. We propose that one of these specific operations involves actively manipulating phonological representations of spoken words in working memory, retrieving various kinds of information about words from the lexicon and transforming the phonological representations of spoken words in various ways depending on the specific task demands.

More detailed analysis of the "Stars" also revealed a very interesting and informative diagnostic pattern of performance on the word recognition tests. The results indicated that the "Stars" organize words in memory and access them using retrieval strategies that are similar to those used by normal-hearing adults and children (Kirk et al., 1995). This conclusion is based on two findings. First, results from the LNT showed that the "Stars" identify "easy" words better than "hard" words. This finding suggests that the "Stars" are recognizing spoken words in the context of other words in memory (Luce & Pisoni, 1998). That is, they are sensitive to the frequency and neighborhood density (i.e., phonetic similarity) of spoken words. This particular response pattern also suggests that the "Stars" have organized words in similarity spaces in lexical memory using equivalence classes. The "Stars" apparently have constructed phonological representations in memory for the sound patterns of words they know and these representations have a particular structure and organization in memory. For the "Stars," spoken words are not just global holistic patterns. Instead, words consist of sequences of sounds with internal structure that are organized in memory in a principled way according to frequency and acoustic-phonetic similarity to facilitate rapid and efficient retrieval.

In addition to showing differences in the recognition of easy and hard words, the "Stars" also displayed sensitivity to word length. That is, they recognized long words better than short words. Again, this particular result indicates that the "Stars" are perceiving sound patterns as words in the context of other words they know. The words they know are organized in a multidimensional space like the lexicons of normal hearing children and adults. Thus, both the easy vs. hard difference and word length effects observed in the LNT indicate that the "Stars" are recognizing words relationally in the context of other words that they have stored representations for in memory (see Luce & Pisoni, 1998).

Although the results from the word recognition tests demonstrate some important similarities in the organization and access of words from lexical memory between the "Stars" and normal-hearing children, there are also several differences in word-recognition performance that should be emphasized here. First, the results of the Minimal Pairs Test indicated that the "Stars" were unable to reliably discriminate differences between pairs of words that differ in place of articulation in a closed-set task until three years post-implant. Even after three years of implant use, the performance of the "Stars" on these very difficult phonetic contrasts is just barely above chance levels of discrimination. This finding demonstrates that the "Stars" may not be able to reliably discriminate and make use of all the fine phonetic detail that is present in speech or encode these small differences in memory. The "Stars" may perceive and recognize words using much broader phonetic categories and may have developed a different set of equivalence classes than normal-hearing children and adults (Koch et al., 1996). The finding that these children can recognize spoken words in isolation does not mean that they recognize these sound patterns in the same way using the same representational specificity and processing operations that normal-hearing children and adults do.

Second, although the performance of the "Stars" on the PBK and LNT tests was exceptionally good in comparison to other deaf children with cochlear implants, the absolute level of performance they achieved on these tests was substantially lower than the scores obtained by normal-hearing three- and four-year old children who routinely display ceiling levels of performance on the same task using the same stimulus materials (Kluck, Pisoni & Kirk, this volume). Thus, although the "Stars" were able to discriminate words using only manner and voicing distinctions, they are apparently able to encode these sound patterns in memory and organize them into a perceptual space using different equivalence classes. Recent analyses suggest that these equivalence classes are broader and less well-differentiated in deaf children with cochlear implants than in normal-hearing children (Meyer, Wright, & Pisoni, 1998). These results are consistent with what we currently know about how the "Stars" discriminate minimal pairs of words and how they recognize spoken words in open-set tests of word recognition such as the PBK and the LNT (Kirk et al., 1995).

Perception and Production

Learning a language involves learning both how to perceive and encode linguistically important differences among spoken words in the language *and* how to articulate these same differences in speech production (Locke, 1993). Differences on the receptive tests of speech and language observed with the "Stars" were also found for measures of speech intelligibility and expressive language. These findings demonstrate transfer of knowledge and overlap between perception and production and suggest that these children have acquired an organized system of knowledge about spoken language, that is, a grammar of the ambient language based on the linguistic input they are exposed to during their first year. The "Stars" appear to be well on their way to learning spoken language and acquiring a grammar from the linguistic input in their environment (Goldin-Meadow & Mylander, 1990). These children showed large and consistent gains in both speech perception and speech intelligibility over the three-year period studied so far. In contrast, the "Control" subjects displayed much poorer performance and even after three years showed only small gains on these tests.

The importance of the correlations between various receptive measures and expressive language as indexed by the speech intelligibility scores needs to be emphasized strongly here because the pattern suggests that whatever the common underlying factor is that shows up repeatedly in these tests, it must be related in some way to the development of an organized system of knowledge that is shared across modalities and is common to both perception and production of spoken language (Prince & Smolensky, 1997). The overall pattern of results strongly suggests that the "Stars" are developing a grammar of the spoken language to which they were exposed (Locke, 1993, 1997). If this is true, then if we look at the correlations with speech intelligibility for the "Stars," we should find converging support from an independent measure of language expressed in a different modality. Whatever kind of learning is going on, it appears to be related to acquiring a shared common system of rules and representations (Pinker, 1991; Seidenberg, 1997), a phonology of the language.

What kind of a mechanism or system might be responsible for the correlations found between speech intelligibility and the language- and speech-related measures, receptive and expressive measures, open-set word recognition measures, and the continued improvements in speech intelligibility and word recognition observed over time with implant use? We suggest that the parallels observed between speech perception and speech intelligibility are consistent with the development of spoken language (Locke, 1993; Jusczyk, 1997). The "Stars" have managed to acquire the phonology of the sound system of the language used in their environment. They have developed an organized linguistic system based on what they can hear through their implant and what they are able to encode and represent in memory in their lexicon. A child who is able to develop a lexicon of the sound patterns of the language displays high correlations with other language measures and perhaps most importantly, shows high correlations with speech intelligibility. The only way this complementary relation between speech perception and production could occur is if the children are able to develop a grammar of the language used in their language-learning environment.

Although the results of the Reynell showed that both the "Stars" and "Controls" are acquiring language, it should be emphasized here that the critical differences between the two groups of subjects may not be due to some global language abilities per se. Instead, we suggest that the differences are much more specific in nature and depend on the processing of spoken words and on that component of spoken language concerned with the development of the lexicon and the abilities used to perceive, encode and retrieve spoken words from memory. The finding that the oral children have much better speech intelligibility scores than the total communication children provides important support for this proposal and is consistent with the hypothesis that the differences we have observed in this study are domain-specific and related to the processing of spoken language in the auditory modality.

Future Directions

At this time, we do not know what factors are responsible for the differences in processing spoken words. Because much of the past research on pediatric cochlear implant users has been concerned with assessment and outcome measures, little attempt has been made to study systematically the nature of the language-learning environment that the child is immersed in after implantation. In future studies, we will need to learn more about the kinds of activities these children are engaged in that use spoken language (Hart & Risley, 1995). What are their parents doing? What kinds of linguistic and metalinguistic activities are they exposed to and involved in on a daily basis? Unfortunately, at the present time, we do not have any measures of how the parents interact with their children at home or the nature of the language-learning environment of these children after implantation (Knutson, Boyd, Reid, Mayne, & Fetrow, 1997). These will be important measures in the future because we know that early sensory experience plays an extremely important role in perceptual and cognitive development and we need to understand more precisely exactly how the language learning environment affects spoken word recognition and language development in these children (Hart & Risley, 1992, 1995; Shatz, 1992).

Knowing more about the reasons that the "Stars" perform so much better than the "Controls" on speech-related tasks suggests several fruitful areas for future research. These new directions focus on measures of *process* rather than traditional measures of audiological *assessment*. Essentially, we want to know how these children process auditory information through their cochlear implants, what kinds of memory codes and neural representations they construct and how they organize and access information from memory (Gilbertson & Kamhi, 1995). These kinds of questions can be addressed directly by studies of perceptual learning, attention, categorization, and especially working memory. For example, what are the characteristics of working memory in children with cochlear implants and how do their memory spans change over time with increases in implant use? Do the "Stars" have longer working memory spans than the "Controls"? What kinds of coding, rehearsal and elaboration strategies do the "Stars" use and how are they different from normal-hearing children and adults? Are the "Stars" able to use context and other information from long-term memory to increase their memory spans? These are important research questions that will need to be studied in the future in order to learn more about the information processing abilities of these children and how these operations are influenced by the impoverished and degraded initial sensory input that the deaf child receives via a cochlear implant. There is now a very sizable literature showing that the ability to perceive and recognize familiar words, learn new words, recall sequences of words and non-words and produce intelligible speech is critically dependent on a small number of perceptual and memory processes that all draw on some aspect of working memory in one way or another (Baddeley et al., 1998; Gupta, 1996; Gupta & MacWhinney, 1997).

Studies of early word learning (Heibeck & Markman, 1987) and research on the organization of the lexicon will also provide new information about how spoken words are encoded and organized in long-term memory and how children with cochlear implants gain access to this information. Normal-hearing children show very rapid word learning, a process often referred to as "fast mapping," which suggests that words can be acquired quickly with very few exposures (Dollaghan, 1985, 1987). This pattern of word learning contrasts markedly with other areas of perceptual learning and development which often show a very slow and gradual process of acquisition. Are the "Stars" children who show "fast mapping" of spoken words like normal-hearing children do? Or, do the "Stars" acquire new words much more slowly over time? We suspect that studies of working memory and early word learning in children with cochlear implants will provide some very important new findings about the underlying basis for the individual differences observed between the "Stars" and the average users on a wide variety of information processing tasks that require processing of spoken words.

Summary and Conclusions

This paper described the results of a correlational analysis of measures of speech perception, comprehension, word recognition, vocabulary, language development and speech intelligibility that was conducted with a small group of exceptionally good cochlear implant users, the so-called "Stars," and a group of "Control" subjects. Our goal was to identify those factors that underlie the exceptionally good performance of the "Stars" and learn why and how they differ from other users. The "Stars" consisted of a group of prelingually deaf children who scored in the top 20% on the PBK test, a very difficult open-set test of spoken word recognition. The "Controls" were another group of prelingually deaf children who scored in the bottom 20% on this test. After the subjects were assigned to these two groups, scores on a variety of other measures of speech and language performance were obtained from an existing database and several analyses were performed to compare differences between the two groups on these test scores. Correlations were also carried out on the scores for each group separately in order to study the relations among the dependent measures.

The results of our analyses revealed several interesting and theoretically important findings. First, in terms of differences in performance between groups, we found that although the "Stars" were consistently better on measures of speech perception, comprehension, word recognition, language development and speech intelligibility than the control group, the groups did not differ from each other on measures of vocabulary knowledge, non-verbal intelligence, visual-motor integration or visual attention. Thus, the "Stars" differed in *selective* ways from the "Controls" and whatever differences were revealed by the other measures, they could not be attributed to a simple global difference in overall performance between the two groups. The "Stars" displayed exceptionally good performance on several other tests of speech and language in addition to the PBK test, which was used as the criterial variable to assign subjects to groups. Thus, whatever skills and abilities the "Stars" have, they are not specific to the PBK test or to other open-set word recognition tests.

Second, the results of the correlational analysis on the test scores of the "Stars" one year post-implant showed a consistent pattern of very strong and highly significant intercorrelations among several of the test measures, particularly tests of spoken word recognition, language development and speech intelligibility, suggesting a common underlying source of variance. The same pattern was not observed for any of the correlations in the control group. Taken together, the results suggest that the "Stars" have acquired language and have developed a grammar based on the linguistic input they were exposed to in their environment. More detailed analysis revealed that the common source of variance found in these analyses was related in some way to the encoding, storage, retrieval and processing of spoken words and the development of phonological and lexical representations for words. Several of the dependent measures were also highly correlated with speech intelligibility scores for the "Stars," suggesting transfer of knowledge and a common representational system between speech perception and production. The two sets of analyses indicate that the exceptionally good performance of the "Stars" appears to be due to their superior abilities to process spoken language, specifically, to perceive, encode and retrieve information about spoken words from lexical memory and use this information in a variety of information processing tasks that require manipulation and transformation of the phonological representations of spoken words.

Finally, several new areas of research on cochlear implants in children were identified that focus on information processing operations including perceptual learning, categorization, attention, and working memory. It was pointed out that traditional audiological assessments of hearing and speech perception performance do not typically measure these types of processing activities and that there is a need for the development of new perceptual and cognitive measures that can account for individual differences and the sensory and perceptual interactions that occur in the language-learning environment during the first year after implantation. More detailed information about the nature of the linguistic interactions between the child and his/her parents and teachers is also needed in order to quantify the nature of the early sensory input the child receives, particularly spoken language inputs during the critical period for language acquisition.

References

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders*. Fourth edition, revised. Washington, DC: Author.

Baddeley, A., Gathercole, S. & Papagno, C. (1998). The phonological loop as a language learning device, *Psychological Review*, 105, 158-173.

Beckman, M.E. and Edwards, J. (in press). Lexical frequency effects on young children's imitative productions. In M. Broe & J. Peirrehumbert (Eds.), *Papers in Laboratory Phonology 5*. Cambridge: Cambridge University Press.

Beery, K.E. (1989). *VMI Developmental test of visual-motor integration*, revised 3rd ed. Cleveland, OH: Modern Circular Press.

Carpenter, P.A., Miyake, A. and Just, M.A. (1994). Working memory constraints in comprehension. In M.A. Gernsbacher (Ed.) *Handbook of Psycholinguistics*, (pp. 1075-1122). San Diego: Academic Press.

Cleary, M. (this volume). Measures of phonological memory span for sounds differing in discriminability using an adaptive-testing procedure: Preliminary findings. *Progress Report on Spoken Language Processing #21*, Indiana University, Department of Psychology, Bloomington, IN.

Dollaghan, C. (1987). Fast mapping in normal and language-impaired children. *Journal of Speech and Hearing Disorders*, 52, 218-222.

Dollaghan, C. (1985). Child meets word: "Fast mapping" in preschool children. *Journal of Speech and Hearing Research*, 28, 449-454.

Dunn, L.M. (1965). *Peabody Picture Vocabulary Test*. Circle Pines, MN: American Guidance Service.

Fryauf-Bertschy, H., Tyler, R.S., Kelsay, D. & Gantz, B.J. (1992). Performance over time of congenitally and postlingually deafened children using a multichannel cochlear implant. *J Speech Hear Res*, 35, 913-920.

Fryauf-Bertschy, H., Tyler, R.S., Kelsay, D., Gantz, B.J. & Woodworth, G.G. (1997). Cochlear implant use by prelingually deafened children: The influences of age at implant and length of device use. *Journal of Speech, Language, and Hearing Research*, 40, 183-199.

Gathercole, S.E. (1995). Is nonword repetition a test of phonological memory or long-term knowledge? It all depends on the nonwords. *Memory & Cognition*, 23, 83-94.

Gathercole, S.E., Willis, C.S., Baddeley, A.D. and Emslie, H. (1994). The children's test of nonword repetition: A test of phonological working memory, *Memory*, 2, 103-127.

Gilbertson, M. & Kamhi, A.G. (1995). Novel word learning in children with hearing impairment. *Journal of Speech and Hearing Research*, 38, 630-642.

Goldin-Meadow, S. & Mylander, C. (1990). Beyond the input given: The child's role in the acquisition of language. *Language*, 66, 323-355.

Gordon, M. (1987). *Instruction manual for the Gordon Diagnostic System*. DeWitt, NY: Gordon Systems.

Gupta, P. (1996). Word learning and verbal short-term memory: A computational account. In G.W. Cottrell (Ed.), *Proceedings of the Eighteenth Annual Meeting of the Cognitive Science Society*, Pp. 189-194. Hillsdale, NJ: LEA.

Gupta, P. & MacWhinney, B. (1997). Vocabulary acquisition and verbal short-term memory: Computational and neural bases. *Brain and Language*, 59, 267-333.

Hart, B. & Risley, T.R. (1995). *Meaningful Differences*. Paul H. Brookes Publishing Co: Baltimore.

Hart, B. & Risley, T.R. (1992). American parenting of language-learning children: Persisting differences in family-child interactions observed in natural home environments. *Developmental Psychology*, 28, 1096-1105.

Haskins, H. (1949). A phonetically balanced test of speech discrimination for children. Unpublished Master's Thesis, Northwestern University, Evanston, IL.

Heibeck, T.H. & Markman, E.M. (1987). Word learning in children: An examination of fast mapping. *Child Development*, 58, 1021-1034.

Jusczyk, P.W. (in press). Bootstrapping from the signal: Some further directions. To appear in J. Weissenborn & B. Hoehle (Eds.), *How to get into language. Approaches to bootstrapping in early language development.* Amsterdam: Benjamins.

Jusczyk, P.W. (1997). *The Discovery of Spoken Language*. Cambridge: MIT Press.

Jusczyk, P.W. & Aslin, R.N. (1995). Infants' detection of sound patterns of words I fluent speech. *Cognitive Psychology, 29*, 1-23.

Kirk, K.I., Pisoni, D.B. & Miyamoto, R.T. (in press). Lexical discrimination by children with cochlear implants: Effects of age at implantation and communication mode. In S. Waltzman & N. Cohen (Eds.) *Proceedings of the Vth International Cochlear Implant Conference*. New York: Thieme Medical Publishers.

Kirk, K.I., Pisoni, D.B. & Osberger, M.J. (1995). Lexical effect on spoken word recognition by pediatric cochlear implant users. *Ear & Hearing*, 16, 470-481.

Kluck, M., Pisoni, D.B. & Kirk, K.I. (this volume). Performance of normal-hearing children on open-set speech perception tests. *Progress Report on Spoken Language Processing #21*, Indiana University, Department of Psychology, Bloomington, IN.

Koch, D.B., Carrell, T.D., Tremblay, K., & Kraus, N. (1996). Perception of synthetic syllables by cochlear-implant users: Relation to other measures of speech perception. Paper presented at the Association for Research in Otolaryngology, Feb. 3-8, 1996. St. Petersburg, FL.

Knutson JF, Boyd RC, Reid JB, Mayne T, Fetrow, R. (1997). Observational assessments of the interaction of implant recipients with family and peers: Preliminary findings. *Otolaryngology–Head and Neck Surgery, 117*, 196-207.

Locke, J.L. (1997). A theory of neurolinguistic development. *Brain & Language*, 58, 265-326.

Locke, J.L. (1993). *The Child's Path to Spoken Language*. Cambridge: Harvard University Press.

Luce, P.A. & Pisoni, D.B. (1998). Recognizing spoken words: The Neighborhood Activation Model. *Ear & Hearing*, 19, 1-36.

Meyer, T.A. & Pisoni, D.B. (submitted). Some computational analyses of the PBK Test: Effects of frequency and lexical density on spoken word recognition. *Ear and Hearing*.

Meyer, T.A., Wright, G.A., & Pisoni, D.B. (1998). Lexical analysis of the errors of pediatric CI users on open-set word recognition tests: A first report. Association for Research in Otolaryngology Abstracts, 21, No. 868.

Miyamoto, R.T., Osberger, M.J., Robbins, A.M., Myres, W.A. & Kessler, K. (1993). Prelingually deafened children's performance with the nucleus multichannel cochlear implant. *Am J Audiol*, 14, 437-445.

Miyamoto, R.T., Osberger, M.J., Todd, S.L., Robbins, A.M., Stroer, B.S., Zimmerman-Phillips, S. & Carney, A.E. (1994). Variables affecting implant performance in children. *Laryngoscope*, 104, 1120-1124.

Monsen, R.B.(1983) The oral speech intelligibility of hearing-impaired talkers. *Journal of Speech and Hearing Disorders*, 48, 286-296.

Newport, E.G. (1990). Maturational constraints on language learning. *Cognitive Science*, 14, 11-28.

Osberger, M.J., Miyamoto, R.T., Zimmerman-Phillips, S., et al. (1991). Independent evaluations of the speech perception abilities of children with the Nucleus 22-channel cochlear implant system. *Ear Hearing*, 12, 66-80.

Osberger, M.J., Robbins, A.M., Todd, S.L. & Riley, A.I. (1994). Speech intelligibility of children with cochlear implants. *Volta Review*, 96, 169-180.

Osberger, M.J., Robbins, A.M., Todd, S.L. & Riley, A.I., Kirk, K.I. & Carney, A.E. (1996). Cochlear implants and tactile aids for children with profound hearing impairment. In Bess, Gravel & Tharpe (Eds.). *Amplification for Children with Auditory Deficits*. Nashville, TN: Bill Wilkerson Center Press, Pp 283-308.

Osberger, M.J., Todd, S.L., Berry, S.W., Robbins, A.M. & Miyamoto, R.T. (1991). Effect of age of onset of deafness on children's speech perception abilities with a cochlear implant. *Ann Otol Rhinol Laryngol*, 100, 883-888.

Pinker, S. (1991). Rules of language, *Science*, 253, 530-535.

Prince, A. and Smolensky, P. (1997). Optimality: From neural networks to universal grammar, *Science*, 275, 1604-1610.

Quittner, A.L., Smith, L.B., Osberger, M.J., Mitchell, T.V. & Katz, D.B. (1994). The impact of audition on the development of visual attention. *Psychological Science*, 5, 347-353.

Reynell, J.K. & Gruber, C.P. (1990). *Reynell Developmental Language Scales: U.S. Edition*. Los Angeles, CA: Western Psychological Services.

Robbins, A.M. (1990). Developing meaningful auditory integration in children with cochlear implants. *The Volta Review*, 361-370.

Robbins, A.M. & Osberger, M.J. (1990). *The Meaningful Use of Speech Scale.* Indianapolis, IN: Indiana University School of Medicine.

Robbins, A.M., Renshaw, J.J., Miyamoto, R.T., Osberger, M.J., & Pope, M.L. (1988). Minimal pairs test. Indianapolis, IN: Indiana University School of Medicine.

Robbins, A., Svirsky, M., Kirk, K. I. (In Press). Children with implants can speak, but can they communicate? *Otolaryngology-Head & Neck Surgery.*

Ruben, R.J. (1992). Language - The outcome measure of pediatric cochlear implantation. Paper presented at the *First European Symposium on Paediatric Cochlear Implantation*, Nottingham, UK, September 24-27, 1992.

Saffran, J.R., Aslin, R.N. & Newport, E.L. (1996). Statistical learning by 8-month-old infants. *Science,* 274, 1926-1928.

Seidenberg, M.S. (1997). Language acquisition and use: Learning and applying probabilistic constraints. *Science,* 275, 1599-1603.

Shatz, C.J. (1992). The developing brain. *Scientific American*, 267, 60-67.

Smith, L.B. & Katz, D.B. (1996). Activity-dependent processes in perceptual and cognitive development. *Perceptual and Cognitive Development.* Pp. 413-445. Academic Press.

Snow, C.E. & Hoefnagel-Hoehle, M. (1978). The critical period for language acquisition: Evidence from second language learning. Child Development, 49, 1114-1118.

Staller, S.J., Pelter, A.L., Brimacombe, J.A., Mecklenberg, D., & Arndt, P. (1991). Pediatric performance with the Nucleus 22-Channel Cochlear Implant System. *American Journal of Otology*, 12, 126-136.

Svirsky, M.A. (1996). Speech production and language development in pediatric cochlear implant users. Paper presented at the 131st meeting of the Acoustical Society of America, Indianapolis.

Waltzman, S.B., Cohen, N.L., Gomolin, R.H., Shapiro, W.H., Ozdaman, S.R. & Hoffman, R.A. (1994). Long-term results of early cochlear implantation in congenitally and prelingually deafened children. *American Journal of Otology*, 15, 9-13.

Watson, C.S. (1991). Auditory perceptual learning and the cochlear implant. *The American Journal of Otology*, 12, 73-.

Zwolan, T.A., Zimmerman-Phillips, S., Asbaugh, C.J., Hieber, S.J, Kileny, P.R. & Telian, S.A. (1997). Cochlear implantation of children with minimal open-set speech recognition skills. *Ear & Hearing*, 18, 240-251.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 21 (1996-1997)
Indiana University

Measures of Phonological Memory Span for Sounds Differing in Discriminability: Some Preliminary Findings[1]

Miranda Cleary[2]

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405

---

[1] This research was supported by NIH NIDCD Research Grant DC00111 and NIDCD Training Grant DC00012 to Indiana University Bloomington.

Measures of Phonological Memory Span for Sounds Differing in Discriminability: Some Preliminary Findings

Abstract. This paper examines the influence of reduced perceptual discriminability on normal-hearing subjects' immediate recall of vowel stimuli edited from natural speech. A newly implemented computerized methodology for obtaining measures of working memory span that permits for convenient change and adjustment of stimulus attributes is described. Among the primary goals of this project was an objective assessment of whether or not this methodology is able to provide useful measures of immediate phonological memory span in individuals, specific to the nature of their perceptual experience. Results indicate that in a group of normal-hearing adults (N=20), the likelihood of correct reproduction was indeed a function of the perceptual discriminability of the vowel stimuli used. Obtained spans were comparable to spans previously reported for more conventional measures. Test-retest correlations for this vowel span are provided from a separate group of 12 subjects to provide some assessment of reliability. Informal testing also indicates that it is reasonable to expect normal-hearing children of about four and half years and older to be able to perform a similar task. It is argued that use of this methodology may be able to help us learn more about the auditory-perceptual and linguistic organization unique to individuals with "non-standard" auditory input experience, such as hearing-impaired listeners, users of cochlear implants, the very young, the elderly, and perhaps non-native speakers of a language.

Introduction

What does a measurement of phonological memory span using an immediate reproduction task allow us to infer about the on-line auditory processing capacity of a given individual? The act of a human listener repeating back a spoken list of words or unfamiliar telephone number exhibits the intertwined effects of sensory limitations, acquired linguistic experience/knowledge, and modality independent cognitive capacities. Numerous factors interact to determine the relative success or failure of this type of reproduction, and current models of memory processes make only a few detailed predictions about how particular acoustic characteristics of the stimuli will affect performance. Moreover, they have little to say about how each individual's unique set of experiences with sound will be reflected in the outcome. For example, very little is known about the effects due to early sensory deprivation or degradation of the acoustic signal on the developmental course of phonological memory, be it long-term phonological memory (mental lexicon organization) or short-term recall ability (see Pisoni, Svirsky, Kirk & Miyamoto, 1997).

The simple hypothesis that was adopted for the purposes of the present study was as follows: shorter phonological immediate memory spans will be associated with lists composed of acoustic stimuli that are perceptually more confusable with other items in the list. This is not a novel hypothesis. Evidence for this outcome has been found using various methodologies (e.g., Conrad, 1964, Baddeley, 1968, Drewnowski & Murdock, 1980), and the effect is robust. However, the results obtained in the past have for the most part not been analyzed for the purposes we have in mind here. Performance on phonological memory span tasks is frequently reported for only very general types of highly familiar stimuli (e.g., digits), and very often the focus is on robust effects across a group of subjects rather than individual patterns of performance. It has been shown, however, that the particular acoustic and phonological characteristics of the stimuli used, both as defined in relation to each other, and to long-term representations of sounds heard previously, have measurable effects on memory span. These effects of what may be roughly termed discriminability and familiarity are interesting for what they reveal about patterns of perceptual organization and cognitive processing ability in groups of individuals who share a similar history of perceptual experience.

The computerized auditory presentation method employed here uses prerecorded acoustic signals with known acoustic attributes and precise timing of presentation. The use of a series of motor responses in a reproduction task rather than a potentially interfering articulatory response may provide a standardizable means of obtaining span measures from individuals whose phonological memory might be under-estimated if an articulated response was the only measure used. These advantages will be discussed in greater detail.

Background on Immediate/Working Memory.

The phonological loop hypothesis. Considerable evidence suggests that specialized circuits exist in the mature human brain for the maintenance of representations that encode acoustic signals, particularly those possessing linguistic and musical structure. This maintenance process has been described in terms of a "phonological loop" (Baddeley, 1986), best thought of as a cyclic "retrieve-and-refresh" process in which re-encoded aspects of the original stimulus energy are fed back and revived at central nervous system levels that can bypass the original sensory input channels. There appear to be two aspects of storage, which, to the extent that order and item recall errors occur independently of each other, can be conceptualized separately: one holds the set of categorized item encodings, and the other, information coding the temporal relation of the items (akin to "loop path"). The re-encoding rehearsal mechanism stimulates the reactivation of items according to the temporal order directions provided by the loop path. The term "phonological loop" is often used to refer to both the storage and rehearsal aspects of phonological memory.

The central executive. The phonological loop mechanism is believed to rely on a more general support structure, which Baddeley and Hitch in their well-known non-quantitative model (1974; Baddeley, 1992), termed the "central executive." Rapid manipulation of limited attentional resources back and forth across different tasks has been found to correlate with activity in the pre-frontal- cortex (e.g., see Goldman-Rakic, 1996). These "centrally executed" operations are thought to coordinate and mediate between modality-specific sub-routines. The executive together with its hypothesized subsystems for the various sensory modalities is often referred to in aggregate as "working memory." In the last two decades, this term has come to be accepted as a useful way of conceptualizing a certain type of brain activity that permits "the temporary holding and manipulation of information during the performance of a range of cognitive tasks such as comprehension, learning, and reasoning" (Baddeley, 1986).

It has long been argued that tasks requiring a novel series of differentiated responses to a recent sequence of events after some imposed length of delay, (whether this delay is occupied by simple cessation of the perceptual experience or, more complexly, an alternate task), must rely on the central executive control circuit and some subset of its "slave subsystems." One of the primary pieces of evidence supporting the notion of a shared central resource is the finding that humans are quite bad at performing multiple such activities concurrently. When the concurrent activities share a particular input modality, the potential for interference effects increases, thus suggesting the existence of modality-specific subsystems. For example, if the stimuli involved are auditory in nature, specialized mechanisms for the maintenance of acoustic stimuli may be activated, and if the stimuli and their context bear sufficient resemblance to an already established linguistic-phonetic identity/category, then the pathways that compose the "phonological loop" can be put in motion. Visual and spatial recall appear to employ similarly specialized networks of connections (e.g., see Jodnides et al., 1996).

Many human abilities, logical discourse and mental arithmetic, for example, are assumed to depend on normal use of these pathways, both for acquisition and for continued function. A number of well-known clinical case studies have been interpreted as providing evidence for or against the "functionally modular" nature of working memory in general, and for the "psychological reality" of its proposed subsystems (e.g., Vallar & Shallice, 1990).

Experimental tasks. A wide variety of laboratory tasks have been devised over the years, each with the intent of exploring a different aspect of working memory processing. Simple digit, letter, and word spans have long been used as basic diagnostic tools for cognitive function. A number of commonly used general intelligence tests (e.g., the WISC-III and WAIS), use digit span tasks that require forwards or reverse order recall of orally produced number strings (Wechsler, 1955, 1991). Rather than general intelligence, per se, for which it is agreed to be a rather poor indicator, performance on the standard forward and backward digit span tasks is argued to be a measure of "freedom from distractibility" as well as auditory short term memory (Wechsler, 1991). The backwards version of digit span is thought to involve an additional "transformational" requirement that tests subjects' ability to manipulate items stored in memory. These tests are usually administered by a live speaker with stimulus presentation times of roughly one item per one or two seconds. Auditorily presented lists, it may be noted, have been generally found to generate slightly longer span measures than visually presented lists of "equivalent" orthographic materials (e.g., see Drewnowski & Murdock, 1980). Other commonly used span tasks include the Knox Cube and Corsi Block measures of spatial memory which require a subject to reproduce a pattern of spatial tapping by the tester on a set of four (Knox) or nine (Corsi) identical blocks arranged on a base (Milner, 1971; Corsi, 1972). Performance on these memory tasks may be tangentially relevant to the study reported here in that the response format used is somewhat similar.

Over the last decade or two, the utility of simple span measures (such as those described above) for explaining more complex cognitive processes has been called into question as a result of reported low correlations between traditional digit span measures, for example, and skills such as reading comprehension (Perfetti & Lesgold, 1977; Daneman & Carpenter, 1980). As a result, researchers have devised other measures that they believe tap into combinations of memory resources crucial to daily cognitive function. The "reading span" measure devised by Daneman and Carpenter (1980), for example, requires subjects to read a list of unrelated sentences, make a true/false judgment after each, and then to recall in order the last word of each of the sentences. Other measures of working memory that also have a history of use include "counting span," in which the subject is asked to count from a sequence of different-colored dot presentations, the number of dots of a particular color, and then recall these totals in sequence (Case, Kurland & Goldberg, 1982), and "digit string span," in which the task is to remember the last digit of a series of digit-strings or results from a series of elementary arithmetic operations (Turner & Engle, 1989). "Matching span" is an arguably easier task which requires the subject to merely indicate whether or not a given list corresponds to one previously seen (Allport, 1984; Martin & Breedin, 1992).

One last task worth mentioning is "non-word repetition." This task involves having the participant repeat back a spoken sequence that may be more or less "word-like" depending on its compliance with the phonotactics of a given language. Responses are generally scored for both phoneme and syllable-level inconsistencies. Non-word repetition appears to demand phonological short-term resources in a manner particularly suited to investigating the influence of long-term lexical representations (or lack thereof) on immediate recall. This measure has recently been used by a number of experimenters to study clinical cases of apparent phonological loop impairment (Papagno & Vallar, 1992; Vallar & Papagno, 1986), and by developmental researchers interested in the emergence of adult-like rehearsal strategies and its relation to vocabulary acquisition (Gathercole & Adams, 1993; Gathercole, Adams, & Hitch, 1994; Gathercole & McCarthy, 1994; Hulme & Roodenrys, 1995). It should be here noted that poor readers, that is, children reading between one and two standard-deviations below their grade level (Rayner & Pollatsek, 1989), generally perform poorly on non-word repetition, more poorly in fact than younger children of matched reading ability, despite there being, in theory, no explicit need to convert from an orthographic code to a phonological one (e.g., see Snowling, 1981). A similar pattern of poor non-word repetition performance has been found in children diagnosed with language impairment (Dollaghan & Campbell, 1997).

Comments on memory span measures. Within each task, a definition of "memory span" is assigned in a rather arbitrary fashion. Traditionally, memory span has often been defined as "the number of items in a list that can be correctly recalled in order, immediately following presentation, half of the time" (Schweickert, 1993). Tests such as the WAIS use an average of the longest three word lists ever recalled during a procedure that tests the subject on three word lists at each increasing list length, with testing halted after two successive failed reproductions. A key point here, however, is that whatever measure of span is used for a single task, the relative performance of different individuals on this task should be preserved across the definitions, as should the relative performance of a single individual with different types of vocabularies. (By my use of the term "vocabularies" here and elsewhere in this paper, I mean to indicate the different types of stimulus items grouped by some common characteristic that are the elements to be recalled.) We will examine the effect of definitional changes of "span" in the data to be presented.

The degree to which these various memory span measures are correlated with each other within individuals has enjoyed a good deal of debate. No consensus has yet been reached about the degree to which individual cognitive differences can be accounted for in terms of general- versus modality/domain-specific capacities. (See Conway & Engle, 1996, and Swanson, 1996, for the general capacity argument; for the domain-specific argument see Daneman & Tardif, 1987 and Jonides et al., 1996.) One general result that has emerged, however, is that simple measures of immediate memory span using verbal vs. visuo-spatial materials, for example, correlate far less well with each other than do more complex working-memory tasks that may use contrastive materials, but share a requirement for attentional manipulation via "central executive function." Moreover, a number of "double-dissociative" clinical case studies seem to indicate that on simple span tasks, modality-specific impairments express themselves, while more involved span measures, presumably involving cross or multi-domain specific capacities, make this dissociation much less apparent. These results are not inconsistent with Baddeley's conception of working memory organization.

Project Motivation.

Our long-term goals parallel those suggested above in the sense that we are interested in what the distribution and development of short-term phonological storage and working memory processing capacity looks like for a population whose source of auditory input is undergoing change/rehabilitation after a profound sensory impairment due to hearing loss. The original motivation for this project can be traced to the finding that open set word recognition performance in prelingually-deafened pediatric users of cochlear implants shows a wide range of individual variability that has yet to be fully accounted for in terms of physiological variables, demographic factors or large-scale psychological measures such as IQ. Neither has there been found a correlation between outcome performance and scores of visual-motor integration or measures of visual attention (Pisoni et al., 1997). The possibility therefore exists that either general capacities such as those subsumed by the "central executive" or domain-specific capacities such as function of the phonological loop may play a role in the outcome differences that are seen in this clinical population. Additionally, it may turn out that the distribution of clinical outcomes parallels in part the distribution of general/domain specific capacities in the general population. The methodology described here aims to provide a practical, efficient, and (in relative terms) "fun" means of obtaining phonological short-term memory span measures from this clinical population. It is first necessary, however, to confirm that normal-hearing adults and children perform as predicted on the task. This preliminary paper details this verification.

Considering that the stimuli used in the present study were a limited set of steady-state vowels and digit-names recorded in isolation, the background to follow concerning the effects of stimulus characteristics on memory span may appear rather excessive. However, because our intention is to test for increasingly more complex effects using variants of this same methodology with different sets of acoustic stimuli more particularly chosen for their acoustic/lexical characteristics, a discussion of the effects we should be able to elicit seems in order. In stating what is already known about the effect of stimulus attributes on immediate recall, this discussion is intended to indicate where our knowledge is still incomplete. The point can also be made that a comprehensive theory that will allow for better qualitative prediction of immediate recall will need to account for the repertory of effects reported in the literature. Anticipating future work involving adults and children with hearing impairments, a selection of related findings regarding phonological working memory specifically in this population will also be presented.

Influence of Stimulus Characteristics on Immediate Recall.

Although Miller's "magical number seven plus or minus two" has become one of the platitudes of the literature (Miller, 1956), it has long been known that particular characteristics of the stimulus items can affect recall performance beyond such bounds. For example, over the years, a number of studies have examined the deleterious effects of lowered signal-to-noise ratios on short-term memory tasks (e.g., Dallett, 1964; Rabbitt, 1968). The effect of noise upon initial identification is the most obvious contributor to this decrement. Although the frequency vs. amplitude spectrum of added noise will determine which featural information will likely be all or partially masked, studies have shown that in the case of consonants with no visual information, place of articulation is the aspect most likely to be lost by lowering the signal-to-noise ratio. The same effect is to some degree achieved by filtering out the higher (and less intense) consonant frequencies. High-pass filtering results in a far less selective decrement, with features such as affrication and voicing also suffering (Miller & Nicely, 1955). Probability of correct identification is ultimately a function of both the number of possible response alternatives and the featural overlap of these alternatives, combined with any knowledge of prior probability of occurrence of such alternatives (and/or features).

Even if the item is correctly identified, the greater difficulty of initial identification alone can affect subsequent recall (Luce, Feutstel, & Pisoni, 1983; Rabbitt, 1968). Because many of the acoustic features used to identify consonants are weaker in intensity and of generally shorter duration than those of vowel portions of the signal, it is not surprising that, regardless of whether syllable structure is CV or VC, recently presented vowels are more likely to be correctly recalled than recently heard consonants from a list of spoken monosyllables, though the size of this effect is greatly dependent on the relative similarity of the consonants also contained in the list (Cole, 1973; Darwin & Baddeley, 1974). The item errors that are made tend to reflect this. Early work by Conrad (1964) for example, suggested that in recalling recorded spoken letter-strings, subjects were more likely to substitute a letter-name containing the same vowel sound as the actually presented letter, than an acoustically dissimilar sounding letter-name or a letter name sharing a consonant sound (although, admittedly, the number of letter-names sharing a vowel sound is far greater than those sharing a consonant sound).

Phonological similarity. It has long been known that when the items in a sequence are phonologically similar to each other, ordered recall is worse than otherwise. The greater the number of phonological "features" the sequence members share, the greater the likelihood of order errors (Wickelgren, 1965). This finding is generally referred to as the "phonological similarity effect" (Conrad & Hull, 1964). The order of an auditorily presented word list consisting of the phonologically similar "mad, man, cad, mat, cap," for example, is less likely to be correctly recalled a few seconds after presentation than another list containing words of comparable familiarity such as "pit, day, cow, sup, bar" that are phonologically unlike each other (Baddeley, 1986).

The above description of the phonological similarity effect should however contain an important proviso: Drewnowski (1980) interestingly showed that the similarity effects associated with the sharing of a common vowel among a read-aloud set of orthographically represented syllabic CV elements could be generated to a certain degree not only when a common vowel was shared among list elements but also when the vowels were acoustically dissimilar but occurred in a well-learned and entirely predictable order. That is, the number of errors made under these two conditions regarding the consonants was not significantly different, suggesting that it is not so much the similarity of the vowels per se that reduces accuracy but rather their lack of distinctive informational value towards encoding the syllable.

Word duration. Although the greater typical duration of a vowel has already been mentioned as a contributing factor towards making this part of a syllable generally more memorable, longer summed duration over all elements of a presented list (of at least X elements) also has a detrimental effect on immediate recall. The word length effect is observed when lists of words that are "equated" in terms of the number of words, syllables, phonemes, and frequency, but that have different average spoken durations are used in a recall task. Lists containing words with shorter average durations are better recalled than lists containing words with longer average durations (Baddeley, Thomson, & Buchanan, 1975). This effect is achieved even when the words are presented orthographically and are controlled for number of printed letters, which implies that a "temporally-extended" phonological re-coding of printed words is maintained.

An "articulatory suppression" task (Baddeley & Hitch, 1974), which requires a subject to engage in or voicelessly plan production of some irrelevant syllable during the presentation or delay period, can, however, obliterate the word length effect, probably through disruption of the rehearsal process. Suppression hides the word length effect in the case of *both* visually and auditorily presented stimuli; it however wipes out the phonological similarity effect only in the case of visually presented stimuli, suggesting that during visual presentation, preoccupation of articulatory planning mechanisms during the suppression task interferes with the usual re-encoding of orthographic stimuli into a phonological representation. Baddeley (1986) has argued that forced articulation during list exposure and delay impairs a mechanism that is responsible for both orthographic re-coding and stored representation "refreshing," but does not prevent the "ready-to-go" phonological representation of an auditory presentation to be registered and stored. Additionally, if a subject is merely exposed to irrelevant speech during the presentation (or retention) interval, the greater the similarity between the distractor speech and the target list items, the worse the interference. The duration of irrelevant items however does not appear to differentially effect recall, suggesting that these items are automatically registered and attended to but not rehearsed to any measurable degree.

A particularly relevant aside on the implications of the general word-length effect on common digit span measures is Ellis and Hennelly's (1980) finding that the average digit span for Welsh-speaking children via the WISC was considerably less than that reported for American children, and that this could conclusively be traced to the longer average vowel duration in the spoken digit names of Welsh. This finding has been replicated across a number of languages and suggests caution when interpreting and comparing span measures across different linguistic populations (Naveh-Benjamin & Ayers, 1986).

Reliability of the phonological similarity and word length effects. Logie, Della Salla, Laiacona, Chalmers and Wynn (1996) provide interesting and important details regarding the types of distributions obtained in the general population for the common word span task, particularly as this concerns the reliability of similarity and word length effects, and the influence of changing strategies on performance. Briefly, Logie et al. found that during a single testing session, 43% of the 251 subjects failed to exhibit one or both of the word length or similarity effects in one or both of the auditory and visual modalities. Although subjects with shorter spans were significantly more likely to fail to display one or more of these effects, there were also some subjects with above average spans that also fell into this category. Although failures to find one or two of the mentioned effects in one or the other of the modalities were not unusual, the incidence of having both effects missing in both modalities was very low.

The important point is this, however: upon a retest, one year later, of 20 subjects who failed to show one or more of these effects, as well as 20 span- and demographically-matched subjects who did demonstrate all four effects during the first session, Logie et al. found extremely low test-retest correlations for the effects, with performance on the first session being very unpredictive of whether or not these effects were observed in the second session. The word-length effect in particular was also noted to be less reliable than the phonological similarity effect. Strategy differences/shifts as ascertained through subject report were found to account for some of the test-retest difference. In their discussion of these results, Logie et al. make several important points about the ramifications of building cognitive theories on group average data. This issue will be revisited in discussing the results presented here.

Effect of voices. In light of the above mentioned variables, it is not surprising that measurable differences in word recognition and serial recall can be found for what are, in orthographic form, "the same lists," depending on the individual attributes of recorded talkers (Hood & Poole, 1980). For example it has been noted in the audiological literature that different recordings of identical words (e.g., the PB-50), as spoken by different talkers for the purposes of standardized assessment, can consistently generate different recognition scores depending on their relative intelligibility (for discussion see Mendel & Danhauer, 1997). Our lab has also found that, given a moderate to fast rate of presentation, immediate recall for word lists with each element recorded from a different speaker is worse than that for sets recorded in a single voice, with the number of errors made being significantly different in early positions of the serial recall curve (Martin, Mullennix, Pisoni, & Summers, 1989). It has also been shown, however, that for these same early items, at slow rates of presentation, talker variability enhances long term recall (Goldinger, Pisoni, & Logan, 1991).

Experience and familiarity effects. As has already been implied, short-term recall does not function independently of long term memory representations. There is ample evidence that familiarity affects speed of identification and rehearsal/sequential access. The highly correlated measure of corpora-frequency has been shown to affect item recall, with lists of high frequency words being easier to recall than lists of low frequency words (other factors being equal) beyond the amount expected purely by greater articulation speed for high frequency words (e.g., as reported by Sumby, 1963; Tehan & Humphreys, 1988). The effect of age of acquisition (which is highly correlated with corpora-frequency) is, however, still under debate, since although it seems to facilitate performance in tasks such as naming (basic level effects), studies such as Roodenrys, Hulme, Alban, Ellis, and Brown (1994) have reported that in word span tasks, no effect is seen when word lists differ only on age of acquisition and are equated for length and frequency. On the other hand, the finding that both children's and adults' memory spans for strings of presumably very high (receptive) frequency function words are worse than those for otherwise balanced strings of content words could perhaps be argued to reflect the influence of age of (productive) acquisition (Humphreys, Lynch, Revelle, & Hall, 1983; Tehan & Humphreys, 1988). In the broadest of terms, however, the generalization can be made that lists of real words are usually better remembered than lists of pseudo-words, which in turn are better remembered than lists of non-words that obey few of the phonotactic constraints of the language (e.g., see Hulme, Maughan & Brown, 1991). Hulme, Roodenrhys, Brown and Mercer (1995) additionally found that subjects whose task was to remember lists of novel words for immediate recall during an initial session demonstrated significantly longer spans upon retest one day later, thus demonstrating the impact of even a limited number of exposures on the interaction between immediate recall and long term memory stores.

Work on the development of non-linguistic expertise has also shown that the degree of familiarity with and/or age of acquisition of the type of strings used in a short term memory task will affect recall (Ericsson & Lehmann, 1996). Also, somewhat relevant to the data presented here (because of the parallel between vowel identity and the character of a complex tone) is Crowder and Surprenant's (1995) finding that trained musicians exhibit a smaller difference in their ability to recall a set of musical tones widely dissimilar on pitch versus a set of more similar tones than a group of non-musicians. That is, both groups found the widely spaced tones from a set of four easier to recall; however, the difference in performance between the two sets was much smaller in the trained subjects. Data from musically-trained subjects who participated in the experiment presented here may also demonstrate this effect.

Combining the acoustic similarity factor with a frequency measure, the prediction would be that highly-familiar/high-frequency words having little acoustic overlap with each other would be remembered most easily. Indeed, using a combined metric of frequency, phonological neighborhood density, and neighborhood frequency (Luce, 1986; Luce & Pisoni, 1998), to create ten-word lists of either non-confusable "easy" (high frequency, low density lexical neighborhoods) or highly confusable "hard" (low frequency, high density neighborhood) words, Goldinger, Pisoni, & Logan (1991) confirmed that "hard" words were less likely to be correctly recalled than "easy" words, though the size of this effect varied over the serial position of the word in the list.

Meaning. Semantic information plays a large role in recall ability during everyday tasks and meaningful chunking has long been known to facilitate memory. Even when such chunking strategies are not consciously adopted, related organizational effects can be observed. For example, lists composed of items all from a single semantic category are better remembered than lists of words drawn from different categories, and lists in which category members appear adjacent to each other are better remembered than lists in which the category items are scattered randomly throughout the list (for further discussion see, e.g., Bower, Clark, Lesgold & Winzenz, 1969; Wetherick, 1975). Semantic factors are often difficult to accurately control in the experimental laboratory, however, in real-world situations, the encoding of semantic relations and elaboration strategies are often among the strongest contributors in determining what is recalled of some input stimulus.

Errors in recall. Up to this point, performance has been described only in general comparative terms. The nature and patterns of the actual errors that are made is, however, important. A computational model, for example, that "showed the word length effect" but made errors in the process that were unlike those likely encountered from a human would leave much to be desired. Item errors (replacement with novel item), transpositions (commonly, swaps of adjacent items [Conrad, 1964]), and omissions often occur, as well as sub-item errors, such as phonological feature deletion or addition. Although it is useful to separate out item errors from order errors, at supra-span levels, these types of errors tend to come hand in hand—

that is, in normal subjects, it is rare to find an individual who makes only order errors beyond their span length, or only errors that are sub-item featural errors.

The types of item and order errors that are typically made for word and syllable lists have been examined by a number of investigators. Drewnowski and Murdock (1980), for example, using a staircase algorithm similar to that employed here, obtained span measures for lists of words, the elements of which were selected randomly with replacement from a large pool of 1008 disyllabic words selected for their heterogeneity of phonological composition. They found that the characteristics most likely to be correctly recalled were the identity of stressed vowels and the overall syllabic stress pattern. In the same vein, Treiman and Danis (1988), showed that the syllable structure of the materials used in a short term memory task affected recall performance. Specifically, they demonstrated the relatively fragile association between the syllable onset and the vowel nucleus, and confirmed that the likelihood of errors at the vowel post-vocalic consonant divide can be influenced by the manner of articulation of the consonant (e.g., liquid vs. obstruent).

The nature of the primary mechanism responsible for recall errors has been an interesting topic of debate for many years. Three relatively independent contributors have been suggested: the interference/overwriting of one trace by another, the passive decay of a trace over time, and the selective encoding of a less than sufficient number of features at time of stimulation (Haber, 1969).

Errors of perception vs. memory? When attempting to model and theorize about human memory processes, researchers sometimes fall into the habit of treating perceptual error and memory error as somehow separable. This often conveniently permits them to remove from consideration large sets of error phenomena "attributable to perceptual confusion" when comparing a model's recall performance with human function. This is, clearly, something of a mistake. A subject's auditory percepts are usually accessible only through subject response "after the fact," and when subject report is relied upon, memory will necessarily play a role. Moreover, memory at various time scales is influenced by attributes of the acoustic signal that go beyond simple intelligibility at moment of presentation. As the pioneers of cognitive psychology clearly understood, to be useful, an account of memory needs to incorporate facts about perception (Haber, 1969; Neisser, 1967).

Development. How these stimulus characteristic effects emerge during development is a matter of considerable interest. The word-length effect for spoken words has been observed in pre-schoolers (Hitch & Halliday, 1983). Hulme (1984) has also demonstrated that phonological similarity effects play an increasingly greater role in error generation as a child gets older and their vocabulary grows in size. (For a related discussion terms of immature organization of the lexicon and the relatively low density of phonological similarity neighborhoods, see Logan, 1992.) Rehearsal strategies related to short term recall are also thought to change over the course of maturation, and thus may be responsible for some of the differences observed in the memory span performance of children versus adults.

The emergence in normal-hearing children of the phonological short-term memory effects described here is far from completely understood. Memory data from certain special populations, specifically from hearing-impaired children, may aid in understanding the process, especially in the case of phonological similarity effects. We therefore now move to examine the long-term impact of auditory deprivation on memory for sound, specifically, speech sounds.

The Impact of Impaired Peripheral Sensory Input.

A great deal of perceptual input has been taken for granted up to this point. Although visual information plays a role in normal speech perception, primary reliance is on the available acoustic signal. Once this signal is compromised by peripheral loss, the realization of the recall effects discussed above can potentially be very different. The discussion presented here cannot do justice to the subtle differences that distinguish different patterns and etiologies of hearing loss and the various counter-measures that may be adopted as a result. Some very rough generalizations regarding the impact of such events on memory for sound sequences will however be attempted.

Post-lingual deafening. Hearing impairment that has ensued well after the period of normal language acquisition and led to use of a hearing aid device, can cause the received signal to be altered in a manner which may make certain sounds more confusable than previously for the hearer. If there is a selective attenuation of particular frequencies typically found in speech, the phonological similarity space for particular phonemes will likely be affected. The pattern of phonological similarity errors may reflect this. The degree to which an experienced hearer can "correct" for this situation by using his or her past knowledge of possible sound combinations for unconscious pattern completion or conscious inference is probably considerable. (See Watson, 1991, for discussion of the learning process that may take place.)

It is not clear what, if any, reorganization of long-term lexical representations might occur over years as a result of a restructured phonological similarity space. Prolonged post-lingual deafness has, however, been shown over time to affect the details of speech articulation, in that, without the usual feedback of their own voices, speakers have increasing difficulty modulating their own voice for the needs of the listener, and also tend to drop articulatory details that they formerly used (Bess & Humes, 1990). With regard to performance on short-term memory tasks, it is unclear whether the mechanisms used in rehearsal would also be affected given this situation, as these are thought to overlap with those used in articulatory planning.

With impaired hearing, and even in the case of profound post-lingual deafness without any residual or rehabilitated hearing, the phonological loop may continue to function relatively normally in situations of reading orthographically presented linguistic stimuli that were likely to have formerly undergone routine phonological analysis. The phonological loop may also be able to use the input gathered during lip-reading via conversion of visually perceived articulation gestures into some kind of phonological code. (The term "phonological" has more recently been used to refer to sub-morphemic units that are simply linguistic primitives, not necessarily sound-based on the periphery, for example, in describing units of sign languages.)

Pre-lingual deafness. The situation of congenitally or pre-lingually deafened individuals is, on the surface, very different. It seems reasonable to predict that pre-lingually deafened individuals who use primarily non-oral methods of communication would demonstrate few effects of acoustic similarity on recall. Contrary to this expectation, however, even native users of sign language have been shown to display short-term memory errors for linguistic material that appear to partially reflect phonetic similarities in the spoken language of their environment. Although Shand's (1982) suggestion that deaf native signers re-encode alphabetic printed words in at least partially terms of manual signs due to the influence of "primary language experience" has not been discarded, the bulk of the literature has suggested that the type of re-encoding Shand described is generally inefficient in practice. There has been some evidence offered that even the profoundly deaf tend adopt some degree of non-manual articulatory encoding for linguistic purposes. This is not to say that the hearing-impaired use strategies strictly analogous to those of hearing persons, (see Marschark and Mayer, 1997, for a better discussion), but rather that an aural-oral-related influence persists alongside manual/formational articulation factors. Hanson (1982), for example, showed evidence in native users of American Sign Language (ASL) for phonetically organized representation via their sensitivity in short-term memory tasks to phonologically (but not orthographically) cued spoken English rhyme and homophony. Such findings have been used to dispute the claim that primary encoding even in native users of ASL is spatial or manual (Hanson & Lichenstein, 1990), though conflicting data (e.g., regarding phonological rhyme sensitivity, Campbell & Wright, 1990) have also been reported.

Much previous research has focused on the question of whether or not lists of silently lip-read stimuli generate the same types of serial recall patterns observed for spoken lists-equivalent recency and suffix effects, for example. Again, contradictory evidence has been reported, (see Campbell, 1987, and Gathercole, 1987, for a review); however, a general consensus was reached in the mid-1980's that work done in the context of serial recall of verbal lists could not be built on the assumption of what was formerly called a "precategorical acoustic store," dependent on acoustic information. This resulted from the finding that serial recall suffix and recency effects for lip-read stimuli resembled recall of auditory stimuli more closely than they did the effects for orthographically presented stimuli.

In individuals raised in an oral communication environment, as might be expected, the oral-aural articulation influence is more pronounced. Conrad (1970) demonstrated that the memory errors made by one group of young, orally-educated, pre-lingually and profoundly deaf male readers in a written-response orthographic consonant letter span task, showed vowel similarity effects (the vowel sounds being contained in the names of the letters) attributable to at least partial oral-articulatory re-encoding of the print. Conrad also found that these particular students were also those who were judged by their teachers as displaying above average speech production ability for those who are profoundly deaf. A later study confirmed that profoundly deaf children differ among themselves in the extent to which they show the effects of articulatory/phonological similarity versus the effects of visual-orthographic similarity in a word span task using lists of printed words (Conrad, 1972).

Dodd, Hobson, Brasher, and Campbell (1983), also found that a group of profoundly pre-lingually deaf adolescent males showed patterns of immediate recall for lip-read digit-names that closely resembled that of hearing peers in that similar recency effects were obtained. Additionally, they observed that a group of orally-educated children with hearing losses of > 70 dB showed a suffix effect for a lip-read suffix but not a non-linguistic tongue-protrusion suffix, in the same manner as normal hearing individuals. Both those students judged as "good" and those judged as "poor" articulators showed these effects. Additionally, Campbell and Wright (1990) found evidence for spoken duration word length effects in young orally-educated, prelingually profoundly deaf students' responses to a pictorial item memory task.

Hearing impaired children also appear to show the same correlation between working memory measures and reading success displayed by normal-hearing children. Daneman, Nemeth, Stainton and Huelsmann (1995), found that the performance of orally educated hearing-impaired (hearing-aid-using) school-age children on a version of the Daneman and Carpenter (1980) reading span task was highly correlated with performance on the Woodcock-Johnson letter-word identification, and prose passage comprehension sub-tests (Woodcock & Johnson, 1977). Three types of working memory span tasks were used in this study; a read-sentences reading span task, a listened-to sentences version of the reading span task, and a visual-shapes array "remember the odd-ball" task. Performance on the reading-span tasks predicted Woodcock-Johnson scores significantly better than did degree of sensory hearing loss which did not predict the Woodcock-Johnson scores well at all. Additionally they noted a small but significant difference between the performance of the hearing-impaired children and matched controls on the spatial working memory span task with the hearing impaired children having slightly longer spans at this task. Otherwise performance between the two groups was quite similar.

The primary interest of this study is measuring the influence of phonological discriminability and distinctiveness on immediate memory. From the evidence discussed above, we can make a strong argument that the relationship between perceptual discriminability and memory in the hearing impaired (or in other contexts, the non-native listener) is not entirely different from the processes used by hearing persons. What is or is not perceptually discriminable, is certainly different, however. What constitutes "phonologically dissimilar" for a normal-hearing individual may not be equally so for an individual, such as a cochlear implant user, for example, with a radically different experience of auditory input. The perceptual similarity space may be "re-organized" with respect to that of the normal-hearing individual/native speaker. The situation becomes even more interesting when one considers the long-term impact of such differences, on the relational properties of the phonological lexical representations that are maintained to recognize words. For example, what constitutes a loosely distributed phonological neighborhood in a normal hearing individual may, in an individual whose auditory apparatus does not permit the same degree of acoustic differentiation, be a much more densely packed neighborhood.

Best's (1994) model of "perceptual assimilation" attempts to predict some of the differences in perceptual organization for speech sound categorization, specifically for non-native speakers of a given language. In the case of individuals with distinctive patterns of hearing loss, prediction of the similarity space poses somewhat different problems for study. There is no guarantee, for example, that the comparison made between the results obtained from the addition of noise to the input of normal hearing individuals is comparable in any way to the experience of anything but a small subset of hearing impaired individuals. Detailed modeling of the non-normal input signal using a series of carefully designed filters assembled specifically to mimic specific cases of hearing loss is one means of at least approximating actual input received in a rational manner (e.g., Humes & Christopherson, 1991).

Overview.

The research reported here was designed to answer several basic questions about the feasibility of using this particular methodology to investigate the types of short-term memory effects discussed in the introduction. This study measured immediate memory span length as a function of vowel-sound discriminability using a memory "reproduction" task. Performance using stimuli drawn from a set of vowel sounds having formant frequencies relatively close to each other was compared with reproduction using a set of vowels whose acoustic characteristics were more dissimilar. Results from two pilot versions of this experiment suggested that this methodology was a workable means of obtaining this general type of data, although some modifications to the methods were necessary. The data obtained using the modified version of the methodology indicated that, as expected, memory for lists composed of the similar vowel set was significantly reduced compared to that for the dissimilar vowel set. Not every subject's individual data reflected this pattern, however. With this particular presentation method, no subject was able to successfully reproduce a list length greater than six for the similar vowels, seven for the dissimilar vowels, and eight for the spoken digit names. Average performance in each of these conditions was considerably less, and this pattern of results remained essentially the same across different definitions of memory span. For point of reference, the average span length for spoken digits reported here is slightly less, but roughly comparable to spans reported in the literature for similar materials.

Method

Participants.

Twenty-two adult subjects, 12 males and 10 females, ranging in age from 18-24 years, participated in the experiment. All but two of the subjects were enrolled in introductory psychology at Indiana University Bloomington and received partial course credit for their participation. The two remaining subjects were volunteers recruited from our laboratory staff. All subjects were native speakers of English. Two were fully bilingual. All reported no known speech or hearing impairments at the time of testing. Two sets of data (from subjects 12 and 14) had to be discarded due to lack of responses. The total $\underline{N} = 20$, 11 males and 9 females.

Materials.

Eight vowel stimuli, Group A: [i], [u], [æ], [ɑ],Group B: [ʌ], [ɛ], [ɪ], [ʊ], each 300 ms in duration, were used in this experiment. Each was edited from a natural speech sample of an American English-speaking male talker producing in isolation (without a carrier phrase), but in several spaced repetitions, his best approximation to steady-state versions of the indicated vowel sounds. The same male speaker was used to generate all stimuli.

Recordings were made in a sound-attenuated single-walled anechoic recording chamber (Industrial Acoustics Company Audiometric Testing Room, Model 402), using a Shure (SM98) microphone. The recordings were digitally sampled on-line at a rate of 22.05 kHz with 16 bit amplitude quantization using a Tucker-Davis Technologies (TDT) System II with an A-to-D converter (DD1), and low-pass filter of 10.4 kHz (anti-aliasing filter, FT5), controlled by an updated version of Dedina's (1987) "Speech Acquisition Program" (Dedina, 1987; Hernández, 1995). The sound files thus generated in .WAV format were checked for possible amplitude clipping and other contaminants. From the several repetitions of the same vowel within the file, a 300 ms segment was selected from the center of a single vowel wave-form that was not the first or last repetition. The first and last 50 ms of each segment were ramped (so as to mimic a natural onset and offset and avoid the perception of "pops" at these locations), and the amplitudes of the entire speech file were normalized to use 80% of available bit space, using the Macintosh application, SoundEdit™ 16 Version 2 (1995) by Macromedia.

Spectral analysis of all stimuli was conducted prior to the experiment using Waves+ from Entropics Corporation running on a Sun Microsystems workstation. Table 1 displays the obtained measurements of F1, F2, F3 and f0 in Hz at approximately 145 ms into the 300 ms segment of each vowel.

The vowel stimuli were next grouped as had been intended, into two sets of four, one set containing the vowels that were maximally dissimilar from each other, and the other, the four remaining vowels. With two levels of phonological similarity as the first independent variable, our dependent variable of interest was within-subject difference in immediate reproduction performance for these two sets of stimuli. The rationale behind this is best illustrated in Figure 1 below, which plots the first formant frequency on the y-axis vs. the difference between the first and second formants on the x-axis, using the frequency values contained in Table 1 converted into values on an auditory frequency Bark scale. Use of this type of scale takes into account the non-linear response of the auditory system at different frequencies. As can be seen in Figure 1, the four vowel stimuli on the outside perimeter of the space are [i], [u], [æ], [ɑ], and neatly within this quadrilateral lie the other "near" vowel stimuli, [ʌ], [ɛ], [ɪ], and [ʊ].

Table 1

Formant and Fundamental Frequencies for Vowel Stimuli

Vowel (as in)	F1(Hz)	F2(Hz)	F3(Hz)	f0(Hz)[a]	@time (ms)
[i] "heed"	263	2206	3071	130	145.9
[ɪ] "hid"	364	1922	2586	131	146.2
[u] "who'd"	287	771	2290	138	145.0
[ʊ] "put"	438	1101	2470	134	145.0
[æ] "pat"	634	1767	2613	124	143.9
[ɛ] "pet"	609	1655	2458	124	146.2
[ɑ] "father"	604	924	2662	127	143.7
[ʌ] "hut"	604	1257	2670	125	143.7

Note. All stimuli were 300 ms in duration. Measurements were taken at the approximate midpoint of the vowel duration as noted in the last column above.

[a] The mean f0 over the eight vowels at the point of measurement = 128.43 Hz, SD = 4.686.

Insert Figure 1 about here

Ten digit-name stimuli *zero* through *nine*, were also recorded as described above, using the same male speaker, except that each digit-name was spoken in the context of the carrier phrase, "_____ is the next number." This intonation pattern was chosen to avoid using word stimuli that would falsely convey, in the context of sequential presentation, the impression of finality, and would better lend themselves to being heard naturally as a list. For the purposes of the present experiment, a subset of four monosyllabic digit names were used: *one, three, five,* and *eight*. Although this choice was somewhat arbitrary, the general aim was to choose names whose measured durations were roughly equal, avoid duplication of the vowel contained, and avoid including digits adjacent in the whole number sequence, (though admittedly however, among the chosen, there remained a sequence of the first three odd numbers). The digit-name stimuli were edited so that by including some small amount of silence at the beginning and end of some words, the edited .WAV files were approximately of equal duration. (Since these were not the stimuli of primary interest, these small durational differences were judged of minor importance.) Amplitudes were again normalized to use 80% of available bit space, using SoundEdit16 Version 2.

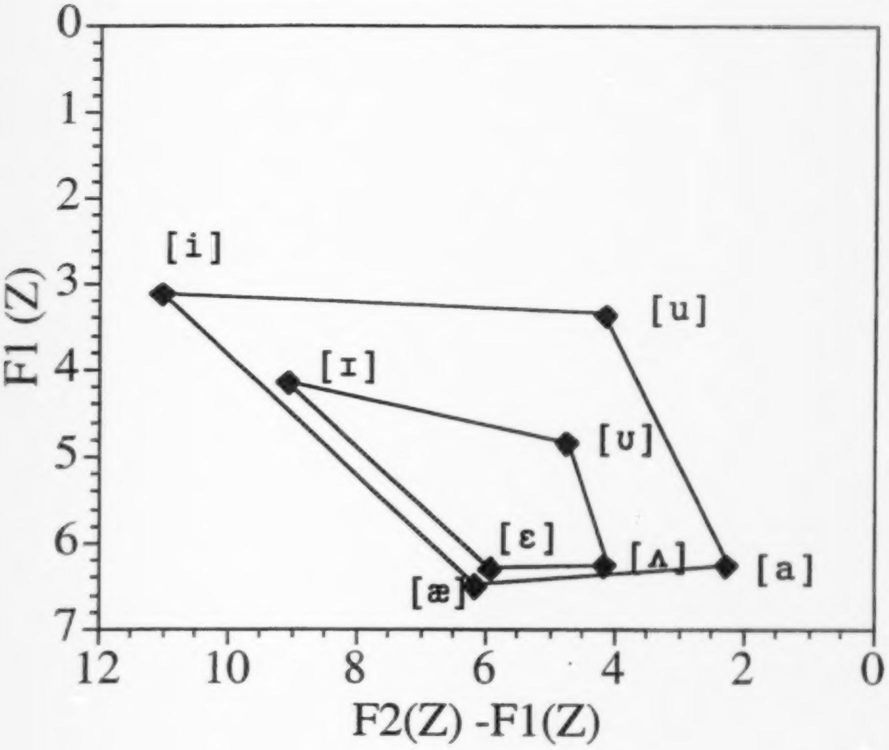


Figure 1. Perceptual space of the vowel stimuli used in this experiment. The difference between F2 and F1 is plotted on the x-axis, the F1 value on the y-axis, both measured in Bark at approximately the halfway point into the 300 ms vowel duration. (B = [(26.81 * f) / (1.96 + f)] – 0.053 where f is in kHz.)

Design and Procedure.

Subjects were run in groups ranging in size from one to five. Each subject was seated within an individual three-quarters-enclosed testing carrel, in front of a computer monitor and a four-button response box, both interfaced to a 133 MHz Pentium Gateway PC with SoundBlaster 16AWE soundboard, timer board, and specialized parallel port. (For details, see Hernández, 1994.) The custom-designed button response boxes employed a diamond-shaped configuration of buttons and location-matched red LEDs. The mapping between stimuli and button position was kept constant across subjects and can be found in Appendix A. Although this relative mapping was kept consistent throughout the experiment, it should be noted that the response boxes are mobile and that subjects were permitted to adjust positioning for purposes of comfort.

Stimuli were played via high-quality headphones (Beyerdynamic, DT 100) at approximately 70 dB SPL (HP voltmeter, max = 0.120 volts, mostly around .060 volts or less = 72 dB SPL). Stimulus presentation and response data collection were controlled by individual computers, each using the stimulus presentation program (written in the computer language C) whose adaptive method-of-adjustment algorithm will be described below.

A within-subjects design was adopted for the purpose of examining individual performance on the different vocabularies. The experiment consisted of three cycles (representing the three levels of our within subject variable), through four different "stages," each cycle using a different "vocabulary" of stimuli, either dissimilar vowels, similar vowels, or spoken digit-names. All subjects heard all vocabularies in the course of the experiment, though not necessarily in the same order. The four stages were termed, 1) "Familiarize," 2) "Learn," 3) "Practice," and 4) "Sequence." Subjects received a set of instructions for the entire experiment prior to the start and were encouraged to refer back to these during inter-stage breaks if they felt the need. These can be found reproduced in Appendix B. Subjects also received a verbal review of instructions and were encouraged to ask questions regarding these before beginning the experiment. During "Familiarize," the participant heard each of the sounds in the current vocabulary played in synch with a light pulse at a consistently matched button location, with approximately 500 ms intervening between each example. Each sound was played twice. Subjects were instructed to merely watch and listen closely. During each of 40 trials in "Learn," the participant heard a single sound through his or her headphones (no light stimuli) and was instructed to press, on each trial, the button that was matched with that particular sound during familiarization. If the correctly matched button was pressed, the matched LED light was displayed and the matched sound played. If an incorrect button was pressed, no sound at all was heard, and a light at the correct location flashed on. This feedback was designed to help the participant learn the correct mapping of sounds to buttons. The subject was made aware of this. During "Practice," the participant again heard a single sound on each of 40 trials and was instructed to press the button matching the sound. If the correct button was pressed the matched sound and light were presented. However, during this stage, if an incorrect response was made, the sound matched to the actual button pressed in response was played and its light flashed on. The participant was instructed not to expect feedback about the nature of the correct response during this stage.

Performance on the fourth and last stage, "Sequence," was the main focus of our investigation. During each of the 25 or fewer trials in this stage, the participant heard a sequence of one or more sounds through his or her headphones. Some stimuli sequences were short, others were longer, as determined by the adaptive algorithm which will be described further. The participant was instructed to reproduce the sequence of sounds by pressing the appropriate buttons to the best of his or her ability, and to wait in order to advance to the next trial.

The experiment was partially counterbalanced for order of vocabularies presented, such that half of the participants (n=10) experienced the dissimilar vowel condition first, followed by the similar vowel condition, followed by the digit-names, whereas the remaining half (n=10) completed the similar, followed by the dissimilar vowels, followed by the digit-names. In both cases, the digit-names were presented last. This was permitted because performance on the digit vocabulary was intended for informal comparison purposes only.

During the stage entitled "Practice" the number of correct vs. incorrect responses was recorded (but not provided to the subject as feedback), and later used to ensure that any subject whose performance on the "Sequence" stage was included in the subsequent analysis was, in fact, able to reliably discriminate between the four stimuli involved.

The design of the stimulus-presentation program used during the "Sequence" stage was a modified version of the adaptive staircase algorithm using a two up/one down rule similar to those described in Levitt (1971). Figure 2 below provides an example of an actual series of 25 trials.

Insert Figure 2 about here

A trial began with the presentation of the sequence to be reproduced, with vocabulary item and order chosen using the computer's random number generator. No element was ever repeated consecutively. Approximately 100 ms of silence was left between each 300 ms vowel presentation, so that items followed each other quickly in succession and participants could easily tell when the last element of the list had been played.

After presenting the list of elements to be reproduced, the computer waited for the subject's response. If 3500 ms passed *without* any button presses, and this was the first or second time this same sequence had been presented on this trial, the sequence was re-presented. However, if at least one button was pressed in response and 3500 ms had passed without a further press, or the sequence to be presented had already been played three times, the subject's responses were checked for correctness.

If the subject's responses were exactly correct and if the immediately preceding sequence (given that there was one) was the same length as the current sequence, (that is, the subject got two in a row correct at a given sequence length), then, on the next list presentation, the number of stimuli presented in a sequence was increased by one. However, if the immediately preceding sequence was a different length than the current one, or this was the very first trial, then on the following presentation, another sequence at the same length was presented. If the response was incorrect in any way, on the next trial a sequence was presented with length one less than the current length, (unless the current length was 1, in which case, the list length was kept at 1). The subject then moved to the next trial and repeated the process, unless a total of twenty-five different sequences had already been presented, or if three full 360-degree oscillations had been recorded around a particular length, in which case, the block of trials ended. (This never occurred.)

As can be seen from the sample in Figure 2, this method of adjusting the list length to be presented will tend to "staircase" up to just above the longest list length a subject is capable of reproducing correctly twice in a row, and then oscillate around this limit. The use of "adaptive" procedures of various types to measure memory span has been shown to yield serial recall functions that are more or less equivalent to the

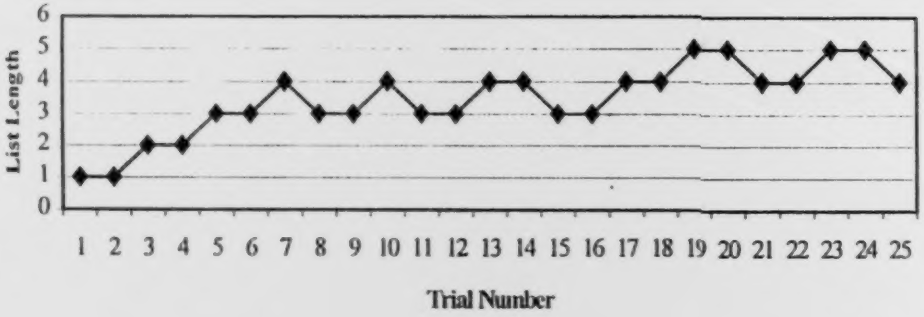


Figure 2. An illustrative sample run of 25 trials using the two up/one down staircase rule under the dissimilar vowel condition. The x-axis represents trial number. The length of the stimulus list presented on that trial is represented on the y-axis.

more commonly used method of constant stimuli (Drewnowski & Murdock, 1980), but avoids potential ceiling effects with particularly adept individuals, and has the advantage of yielding a maximum number of trials precisely around the list lengths that flank an individual's memory span. The two up/one down rule used yields, on average, an approximately 80% correct response rate.

Data files of a standardized format were generated individually for each subject. Analysis of results was conducted using a combination of Microsoft Excel, SPSS, and Statview, all for the Windows environment. Some hand-scoring was also necessary in the early stages of the data analysis.

Subjects' "span" for each type of stimulus vocabulary was determined according to three different criteria and the results independently compared. The first of these, the least conservative in nature, was a definition of memory span as "the longest list length a subject was able to correctly reproduce at least once." The second criterion used was a calculation of memory span as "the longest list length a subject was able to correctly reproduce on least 50% of all trials given at that length." The last and most conservative criterion for memory span was "the longest list length using a particular vocabulary that the subject was able to correctly reproduce on 100% of all trials at that length."

Note that due to the adaptive algorithm used, not all subjects experienced the same number of trials at a given list length. However, also due to the algorithm, it was the case that all subjects received proportionally more trials at lengths that were neither impossible, nor absurdly easy for them to reproduce.

Results and Discussion

The data for any subject demonstrating a clear inability (defined as fewer than 30 trials correct out of 40 total per vocabulary) to reliably make the auditory discrimination necessary to learn the correct-button mapping in the 40 practice trials provided, were discarded in the analysis. Chance would be 10 correct. The data from two subjects were removed from the final analysis as a result, leading to a total N of 20.

Differences across the different types of stimulus vocabularies were obtained in the "Practice" results. Specifically, the distribution of "Practice" scores was wider for the acoustically similar ("near") vowels than for the dissimilar ("far") vowels, which in turn, was larger than that obtained for the digit-names. Table 2 shows the means and standard deviations for these "Practice" scores. This difference will account for some proportion of the difference in sequence recall as a function of vocabulary.

Table 2

Number of Correct Responses During 40 "Practice" Trials as a Function of Vocabulary

	Stimulus Vocabulary		
	Digit-names	Dissimilar Vowels	Similar Vowels
M	39.89	39.33	37.56
SD	0.31	0.88	2.50

Figure 3 shows, for the vowel conditions, the number of trials at each list length that resulted from the list-length adjustment rule used. Note that the difference in shape of the two distributions gives some indication as to the relative ease of the two vowel vocabularies. The total number of trials represented in this graph sums across all subjects to 500 trials of each vocabulary type.

Insert Figure 3 about here

A comparison of the experimental groups provided some evidence for order effects related to vocabulary type. Specifically, subjects that completed the dissimilar vowel condition, followed by the similar vowel condition, generally did better on all conditions than did subjects who completed the similar condition first. (See Figure 4.) As computed using the most conservative definition of span, this effect was significant at the $p < .05$ level for the similar vowel set vocabulary, and although the effect was not significant for the other two vocabularies (note the large variances), mean spans for the other two vocabularies in the group that first completed the more difficult similar vowel condition were lower than those of the group that completed the dissimilar vowel condition. Differences in performance at list lengths 2 and 3 accounted for most of these span differences between the groups. Note that since subjects who did the harder condition ("Similar Vowels") before the easier condition ("Dissimilar Vowels") did not seem to benefit from the experience of having completed the previous condition enough to make their performance on the easier condition comparable to that of the group which completed this condition first, this may not be a simple practice effect. Proactive interference may be responsible in part for these differences, or subjects who completed the harder condition first may have become discouraged early on and subsequently failed to engage the task with the same effort as the other group. Since there was no evidence for an interaction, however (that is, spans in the Similar/Dissimilar/Digit-names group were uniformly lower for every vocabulary), for the remainder of the analysis, scores from both groups of subjects have been combined. Analyses of individual groups separately revealed the same direction of span differences across vocabulary as are presented here, though in some cases, specifically in the group that demonstrated overall longer spans, the difference in means between the two vowel conditions was smaller and non-significant in paired t-tests at the .05 level, though it lay in the expected direction.

Insert Figure 4 about here

Again, note that both groups completed the digit-name condition last. The higher span means obtained for this vocabulary likely result from both practice effects and the influence of high familiarity with these phonological forms. Although this vocabulary is included for interest's sake in the graphs to follow, since its presentation was not balanced across groups, no serious conclusions should be drawn from the data for this vocabulary.

Percent correct reproduction at each list length was computed from each subject's total number correct at each list length divided by total number of trials presented to the subject at that length. The average of these scores across subjects is shown in Figure 5 as a function of list length and vocabulary type. Note that the number of total trials from which these means are calculated, changes as a function of list length as shown previously in Figure 3. Statistics will not be presented for this mean percent correct directly but rather for the span scores as determined by the three different criteria previously described. The relative difficulty of the three different vocabularies is evident in Figure 5 at every list length.

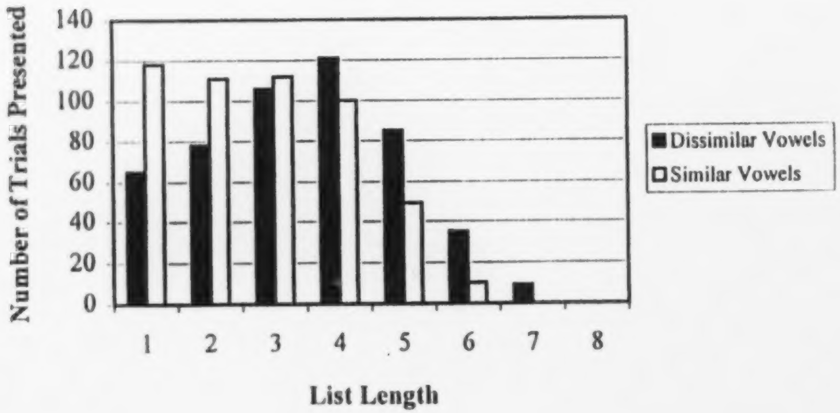
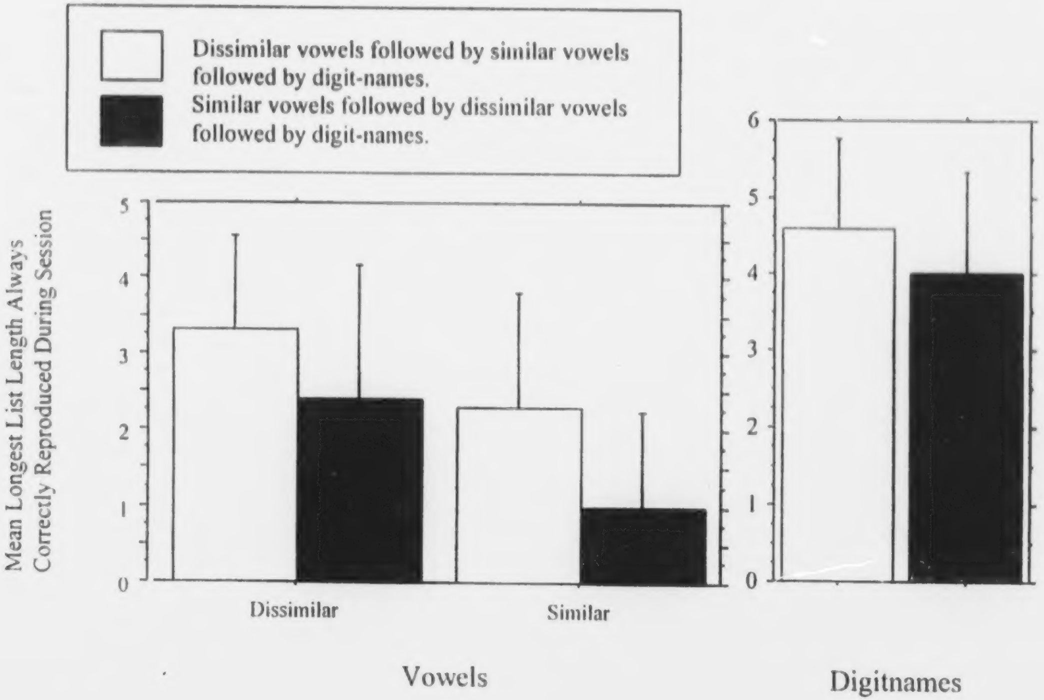


Figure 3. Across all subjects (N = 20), the number of lists presented at each list length. Dark bars show the distribution of lists for the dissimilar vowel condition. Light bars display the distribution of lists under the similar vowel condition.



Note. Bars indicate 1 SD

Figure 4. Influence of condition order on mean span. Span is defined as the longest list length correctly reproduced on all trials of that length for each vocabulary. Unfilled bars represent condition in which participants experienced the dissimilar vowel vocabulary, followed by the similar vowel vocabulary, followed by the digit-name vocabulary. Filled bars represent condition in which participants experienced the similar vowel vocabulary first, followed by the dissimilar vowel and digit-name vocabularies.

Particularly striking is the difference in how quickly performance falls off as list length increases. For the digit-names, performance remains quite high right through until list length 4 and then drops off relatively quickly, although even at list length 7, 34% of the digit-name lists at that length were correctly recalled by subjects. The vowel vocabularies show a more gradual decline, with decrements in performance clearly evident already at list lengths 2 and 3.

Insert Figure 5 about here

Figure 6 below shows span defined as the longest list length reproduced at least once, as a function of vocabulary. Note again that this is the least conservative definition of span used in the analysis. The mean spans and standard deviations for each vocabulary are listed in Table 3. A repeated-measures analysis of variance between all three vocabulary types yielded an $F(2,38) = 22.562$, $p < .0001$, indicating a significant effect of vocabulary type, and a paired t-test comparing only the dissimilar vs. similar vowel conditions yielded a $t(19) = 2.268$, $p < .05$ ($p = .0352$) with spans for the dissimilar vowel condition being significantly higher than for the similar vowels.

Table 3 provides Pearson's product-moment correlations between the vocabularies across all subjects. This measure will tend to reflect the degree to which individual subjects maintain their performance relative to the group across the different vocabularies. Spans for the dissimilar vowel set showed a moderately large positive correlation with spans obtained for the spoken digit-names. Correlations between spans for the two vowel conditions and between the similar vowel condition and the digit-name condition were small and unremarkable. It may be noted that these generalizations held, though correlations were somewhat smaller, when the counter-balanced groups were analyzed separately.

Insert Figure 6 about here

Table 3

Spans and Correlations Between Spans Obtained According to Criterion of Longest List Length Correctly Reproduced At Least Once

	StimulusVocabulary		
		Vowels	
	Digits	Dissimilar	Similar
M	6.15	5.05	4.35
SD (sample)	0.85	1.02	1.19
Mdn	6	5	5
			r
Span for Digits vs. Dissimilar Vowels			.68
Span for Digits vs. Similar Vowels			.19
Span for. Dissimilar Vowels vs. Similar Vowels			.27

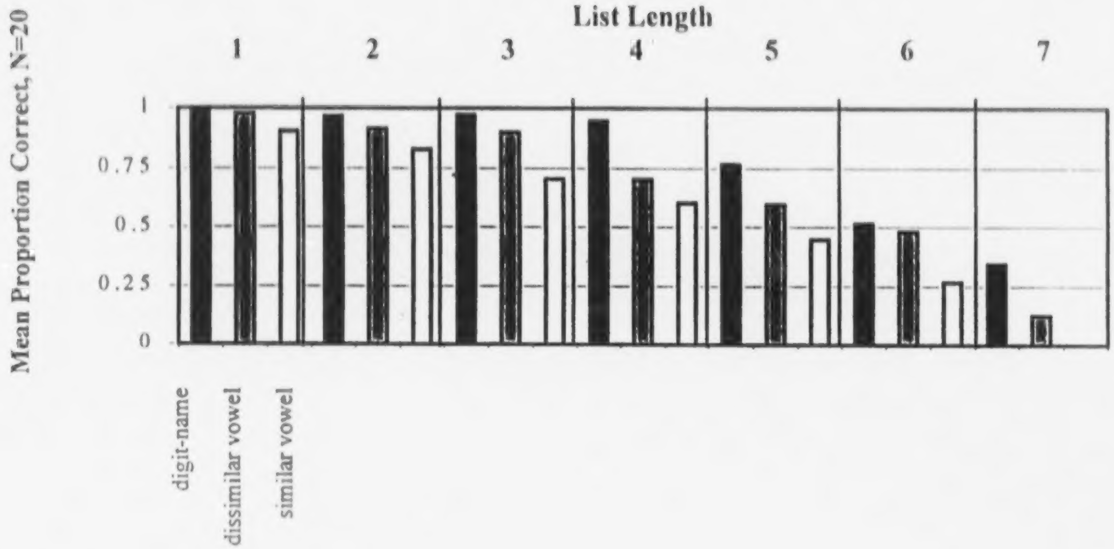


Figure 5. Mean proportion perfect reproduction as a function of list length and vocabulary condition.

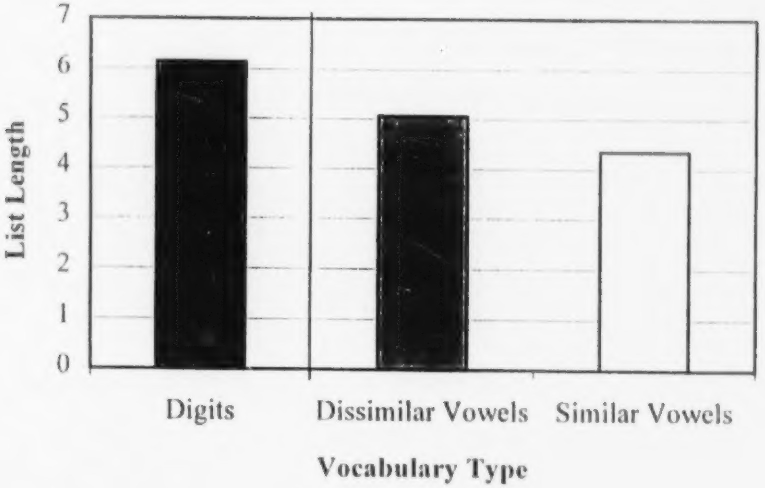


Figure 6. Longest list length correctly reproduced at least once, means across subjects as a function of vocabulary.

For the spans obtained from using a criterion of the longest list length produced at least 50% of the time, a repeated-measures analysis of variance between all three vocabulary types yielded an F(2,38) = 19.033, p < .0001, again showing a significant effect of vocabulary type, while a paired t-test comparing only the dissimilar vs. similar vowel conditions yielded a t(19) = 2.545, p <.05 (p=.0198). (See Figure 7.) A moderately large positive correlation was found between spans for the dissimilar vowels and the digit-names using this criterion as well. That is to say, spans on the digit-name and dissimilar vowel vocabularies were rather well able to predict each other, likely due to the relative perceptual discriminability of the dissimilar vowels approaching that of the digit-names.

Insert Figure 7 about here

Table 4

Spans and Correlations Obtained According to Criterion of Longest List Length Correctly Reproduced on at Least 50% of the Trials Presented at that List Length

	StimulusVocabulary		
		Vowels	
	Digits	Dissimilar	Similar
M	5.80	4.85	3.90
SD (sample)	0.87	1.19	1.34
Mdn	6	5	4
			r
Span for Digits vs. Dissimilar Vowels			.69
Span for Digits vs. Similar Vowels			.20
Span for. Dissimilar Vowels vs. Similar Vowels			.18

Lastly, under the most conservative definition of span as the longest list length a subject was able to reproduce correctly on 100% of all trials at that length, a repeated-measures analysis of variance between all three vocabulary types again showed a significant effect of vocabulary type with an F(2,38) = 29.805, p < .0001, while a paired t-test comparing only the dissimilar vs. similar vowel conditions yielded a t(19) = 3.479, p <.01 (p=.0025).

It should be noted, however, that the use of this strict criterion tended to reduce spans quite drastically for some subjects in one or more of the conditions, and not necessarily for the condition in which the subject had the lowest span according to the two previous criteria. Larger variances, in proportion to the mean spans were an inevitable result, as can be seen in Table 5. This contributed to rather erratic and difficult to interpret non-corresponding patterns of correlations within the counter-balanced groups.

Insert Figure 8 about here

BEST COPY AVAILABLE

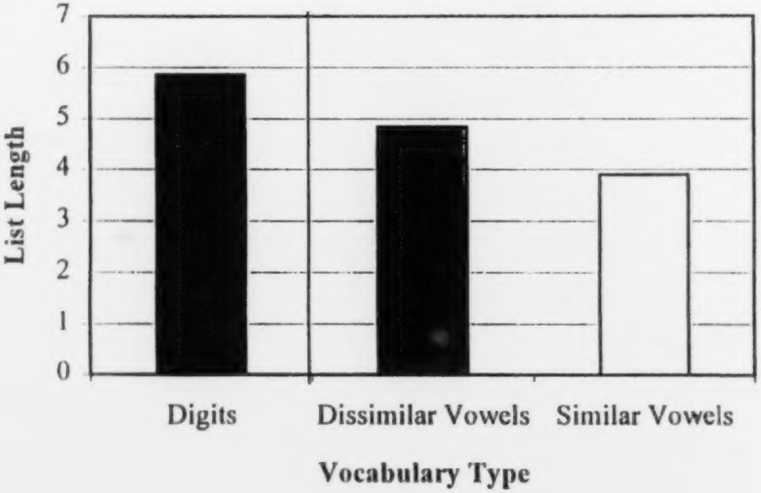


Figure 7. Longest list length correctly reproduced at least 50% of the time for all trials at that length, means across subjects as a function of vocabulary.

M. CLEARY

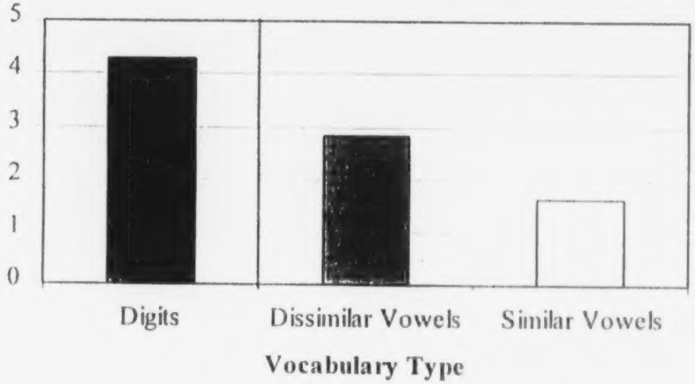


Figure 8. Longest list length correctly reproduced on all trials at that length, means across subjects as a function of vocabulary.

Table 5

Spans and Correlations Obtained According to Criterion of Longest List Length Correctly Reproduced on at Least 100% of the Trials Presented at that List Length

	StimulusVocabulary		
		Vowels	
	Digits	Dissimilar	Similar
M	4.30	2.85	1.65
SD (sample)	1.23	1.53	1.46
Mdn	4	3	2
			r
Span for Digits vs. Dissimilar Vowels			.34
Span for Digits vs. Similar Vowels			.48
Span for. Dissimilar Vowels vs. Similar Vowels			.49

Note. These correlations may not be as meaningful as those given for the previous criteria of span. See text for discussion.

Some comment seems in order regarding the nature of the errors obtained using the different vocabularies, despite the fact that only four stimuli were used as possible responses and that the algorithm used pre-ordained that only about 20% of all responses would contain errors. Nevertheless, regarding these approximately 300+ erroneous responses, a few observations can perhaps be made:

First, very few erroneous responses were made in which all items were correct and only order was incorrect. (7 for the digit-names, 22 for the dissimilar vowels, 8 for the similar vowels.) This may be partly a result of the instructions received by the subjects (see Appendix B) which can be interpreted as encouraging subjects to correct errors they notice and to continue pressing the remainder of the sequence, thus leading to responses longer in length than the sequence presented. Of the small number of pure order errors, of those which involved only the reversal of two consecutive items, in the dissimilar vowel condition, 8 of the total 16 such errors involved reversing the consecutive order of [æ], [ɑ]—vowels with similar F1s. In the similar vowel condition, 4 of 7 total errors defined in the same manner involved reversing the order of [ʌ], and [ɛ], also the two stimuli with the closest F1s in the set. No other error patterns were evident.

An additional result, which will be expanded upon in the general discussion, concerns individual patterns of performance. The group data clearly shows a pattern of results that seems to lead to the expectation that in an individual, the memory span calculated for the perceptually more discriminable stimuli will be greater than that obtained for the perceptually more similar set of stimuli. Strictly speaking, however, this expectation was not reliably met on the individual level, in the sense that, via any of the three ways span was defined here, no more than half the subjects (and not necessarily the same ones, depending on the criterion), met all parts of the informal prediction of span for digit-names > span for dissimilar vowels > span for similar vowels. Most subjects showed at least one of the expected inequalities, and comparison of individual performance on only the two vowel conditions, showed that a slim majority of the subjects met this part of the expectation under each definition of span (10/20 for "at least once," 11/20 for "at least 50%" and 13/20 for "100%"). The remaining subjects either performed equally well on both sets of vowels or displayed a longer span for the similar vowel stimuli than for the dissimilar vowels (6/20 equally well, 4/20 opposite for "at least once," 6/20 equally well, 3/20 opposite for "at least 50%," and 6/20 equally well, 1/20 opposite for "100%").

Anecdotally, it might also be noted that two subjects who revealed considerable musical training appeared to find the vowel conditions considerably easier than other participants. Both of these participants mentioned utilizing the pitch intervals between sequence items in order to aid their recall. As noted in the methods section, f0 values did range between 124 to 138 hz across the dissimilar vowels and from 124 to 134 hertz across the similar vowels. This was true despite efforts during recording to keep the f0 more or less constant—vowels tend to have some inherent f0 differences in production. Specifically, the vowels with higher F1s ([i], [u], [ɪ], [ʊ]) displayed higher measured f0 values. Spans appeared to be longer for the musical subjects in the vowel conditions than the average, and the difference in performance between the two vowel sets seemed to be less, though this difference has yet to be assessed formally.

Preliminary Test-Retest Reliability for Vowel Span Measures.

A separate group of 16 adult subjects was recruited in order to examine the degree to which individual performance remained stable across two sessions scheduled approximately one week apart. These subjects received course credit for their participation. Three subjects were unable to complete the task, and the data from one subject who reported using a musical pitch-related strategy and demonstrated a markedly different pattern of results was examined separately. Of the remaining 12 subjects, all completed a first session in exactly the same manner as the subjects described above, with half completing the dissimilar vowel condition first, then the similar vowel condition followed by the digit-names, and the remainder, the similar vowel condition first, then dissimilar vowel condition, then the digit-names. Upon returning approximately one week later (mean = 6.5 days), each subject repeated the vowel condition they had first encountered during the initial session, followed by the digit-name condition, followed by a third condition not relevant to this paper. Prior to the sequence reproduction stage for each vocabulary, subjects underwent the same familiarization and learning stages as they had during the first session; however, the additional testing stage was eliminated as the subjects were by now very familiar with the stimuli and their matched locations.

Although additional subjects will need to be run in order to draw any strong conclusions about the reliability of this measure, preliminary Pearson's correlation coefficients for memory spans defined as the longest list length ever completed correctly at least once were $\underline{r} = 0.442$ for the dissimilar vowel set ($\underline{n}$=6), and $\underline{r} = 0.489$ for the similar vowel set ($\underline{n}$=6). Using the most conservative definition of span, that is, the longest list length always correctly reproduced on all trials at that length, $\underline{r} = -0.120$ for the dissimilar vowels ($\underline{n}$=6), and $\underline{r} = 0.644$ for the similar vowels ($\underline{n}$=6). These results are better illustrated in Figures 9 and 10 and can be interpreted to suggest that individual performance on this task can be expected to vary somewhat from session to session. Using a less conservative span measure however, will assure that this variability will be less than if the more conservative criterion is used, as evidenced by the fact that very few subjects obtained a span score on the second day that was not within one unit of the first score when span was measured as the longest list length correctly reproduced at least once during a session, but that when the more conservative measures was employed, many more subjects generated spans that differed by at least two units on the different days. The scatter-plots in Figures 9 and 10 show that subjects' memory span performance did not necessarily improve between the first and second sessions.

Insert Figures 9 and 10 about here

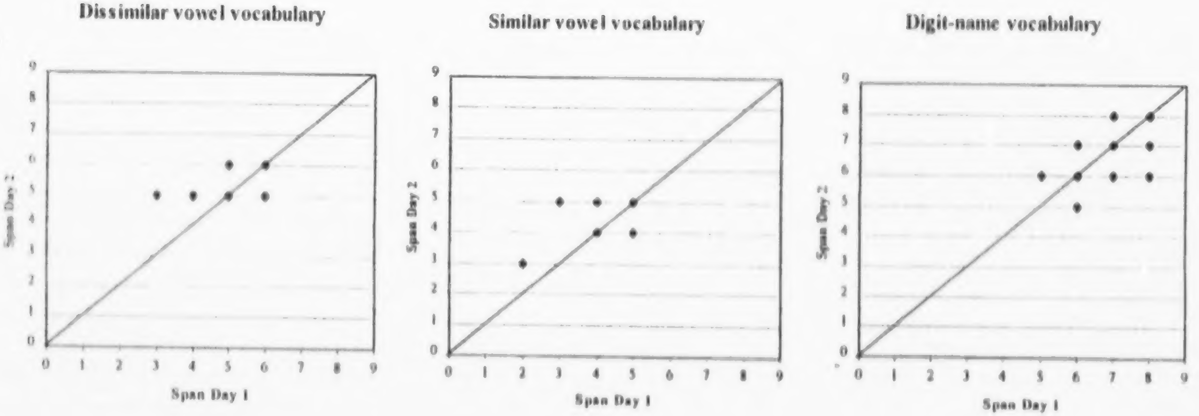
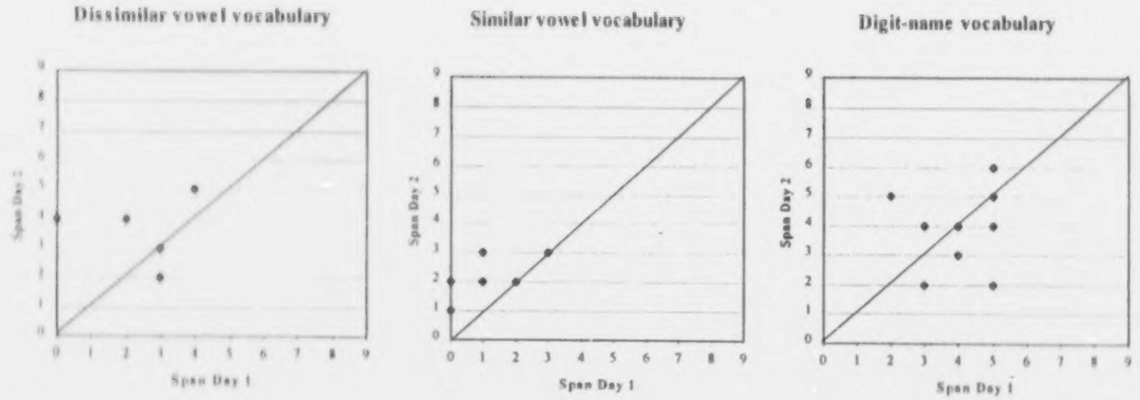


Figure 9. Comparison of individual span scores over 2 sessions using least conservative span measure. Spans obtained on Day 1 are plotted on the x-axis, spans obtained on Day 2, on the y-axis.



Dissimilar vowel vocabulary

Similar vowel vocabulary

Digit-name vocabulary

Figure 10. Comparison of individual span scores over 2 sessions using most conservative span measure. Spans obtained on Day 1 are plotted on the x-axis, spans obtained on Day 2, on the y-axis.

Child Pilot.

An informal pilot study (N=8, ages, 4-7) indicated that the nature of this task is such that it can be easily used with some children as young as four years of age. The apparatus used with the children was an electronically modified version of the well-known Milton Bradley game Simon™, whose four brightly-colored buttons and iridescent casing easily attract the interest of young participants.

Experimenter tact and enthusiasm were important for keeping the child on task. A couple of notes regarding this procedure are, I think, worth noting here for the benefit of other researchers interested in using this methodology. Our initial plan was to provide mildly positive verbal feedback on each trial, regardless of the correctness of the child's response. Some obvious problems arose with this, the first being that many children clearly realized when they had made a mistake on their own and found positive feedback on these trials odd, and even annoying. The issue of reinforcing wrong answers was a problem with the younger children too, as it allowed them to wildly stray from the instructions they received. These problems were partially resolved by having the experimenter adopt a more or less hands-off, neutral-feedback approach towards the child's performance. A "practice" set of presses was also added before each vocabulary type during which each child could experiment with the feel of the buttons, and could be checked as having understood the instructions. A concern also came up of whether the child could take cues from the experimenter's eye-movements or expression, about what the next correct button press would be. This was countered by having the experimenter more or less direct their eyes to a neutral point in space and making an effort to suppress reaction. More to the point, according to anecdotal report and general impression, the child subjects tended not to look at the experimenter between the sequence presentation and their self-initiated response, and if they did, it was usually after finishing their response, with a smile of believing they had correctly reproduced the sequence, a chagrined comment, or an inquiry of when the session would be over. The children were told and seemed to understand that the computer was controlling the presentation of the sequences, and not look to the human experimenter for cues, much in the manner of playing a video or computer game.

Figure 11 below shows the performance of a 6 ½ year-old child using this setup with lists of dissimilar vowels. As can be seen, this particular child's memory span performance is "oscillating" roughly between list lengths 3 and 4, and closer to 3, since at no point was the child able to correctly reproduce in succession two lists at list length 4. This child is clearly able to complete the task requirements, as were most of the other children who participated in this small pilot. We think it will be feasible to use this method to examine increases in immediate memory span during the early elementary school years and the types of strategies children learn to employ.

Insert Figure 11 about here

General Discussion

The results presented above are not surprising in that they replicate the finding that phonological similarity produces a measurable effect on the probability of correct immediate reproduction. To some degree, we can therefore be confident that this method of collecting measures of auditory memory spans is a legitimate way of gathering such data. At the same time, however, an issue has been raised concerning the relative difficulty of the stimulus vocabularies with respect to the order in which they are encountered during a testing session. This commonly encountered problem in individual testing will need to be taken

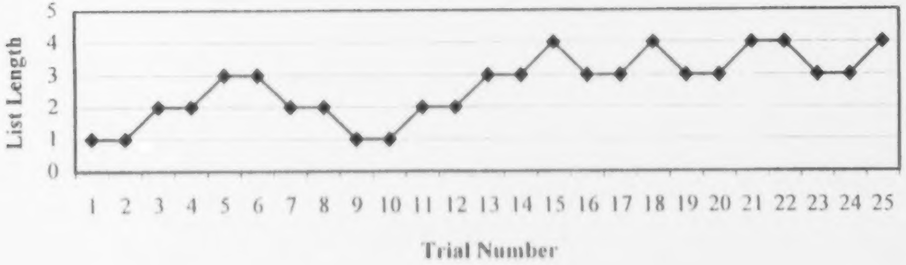


Figure 11. A 6 ½ year-old child's performance on a similar reproduction task with the addition of light sequence cues. An illustrative sample run of 25 trials using the two up/one down staircase rule under the dissimilar vowel condition.

into account when this measure is used. We have also begun to answer the question of how effective and/or reliable this type of measurement may be at accurately indicating individual differences in immediate phonological reproduction ability.

This research represents a preliminary step in our search for an experimental methodology to measure phonological memory span efficiently and reliably using different types of acoustic stimuli and timing manipulations with a wide variety of subject populations. Because we are interested long term in both the perceptual and developmental aspects of phonological short term memory, our goal is to develop a suite of standardized testing procedures involving tasks which young children are capable of performing, yet which will generate results not susceptible to ceiling-effects with more mature subjects. We think it likely that some form of the current methodology will find its way into such a toolbox, but some further modification may be in order.

In continuing this line of research, we will attempt to present further evidence that some of the mechanisms utilized for speech perception are also used in a phonological reproduction task which more transparently draws upon immediate/working memory capacity. In this preliminary study we have shown that the reliable finding of perceptual discriminability having a significant impact on probability of correct reproduction of auditory sequences is replicated using this methodology with normal-hearing adults. Using a closed-set response format, within-subject and between-subject differences in reproductive accuracy have been examined as a function of stimulus type and compared to what we know about the average perceptual distinctiveness of the items involved.

Some proportion of the relative ease of reproducing sound sequences such as those used here can be predicted by an "on the average" estimate of perceptual similarity. Individual differences in performance involve other considerations however. Differences specific to the individual in central processing, and in peripheral perceptual abilities, both of which will reflect experience-dependent factors, can be expected to be among the contributors to this variability.

These are not, however, the only possible influences, as must be kept in mind when trying to use individual performance on a simple memory span task to infer something about the processing capabilities of an individual. It is clear from the data presented here that much care will be required in interpreting a specific set of results if one is attempting to draw conclusions on a case-by-case basis. Simply because an individual generates memory reproduction scores during a particular session that run counter to the patterns predicted by group average discriminability rankings does not necessarily imply that that individual's perceptual and phonological memory retrieval processes are different than expected. Factors such as anxiety and motor coordination can be anticipated to affect scores, for example, although this study does not specifically examine such influences. Experience with the task, however, has been shown to have a large effect on performance, though exactly how this operates is a matter for further investigation. Simple spoken digit span tasks are reported in the literature as generating practice effects that are statistically significant but "negligible" (Lezak, 1995). For less familiar vocabularies, however, there are suggestions that the learning rate responsible for the increase in performance over time with the same vocabulary is steeper and the practice effect larger. The attempt made here to assess the reliability of "span" as defined in any of the three ways described previously, illustrates, I believe, that the expectation of generating a particular set value of "span" for a given subject cannot and should not be based on a single previous measurement and or session. An individual will generate different span scores from session to session and at best, some average of these may be able to give us some ability to predict reasonably well his or her performance on a subsequent session. Decidedly mediocre correlation values were obtained for the small group of subjects that returned for the repeat session. We are currently exploring the possibility that the integer span units that were used for scoring may have collapsed the scores to the point of insensitivity. We note here that test-retest reliability coefficients spanning a range of .66 to .89 have been reported by other researchers for spoken digit-names (values cited in Lezak, 1995).

Even a good deal of practice at identifying the individual elements of an acoustically similar set of elements will not necessarily translate into optimal performance when sequences of these elements are encountered. Practice with the sequence format response, even with another vocabulary seems to boost performance with highly confusable stimuli. Interestingly, however, difficulty with a confusable vocabulary during the early part of this type experimental session appears to translate into diminished performance on the easier vocabularies later in the session relative to a control group. This is a shaky observation, however, and the inclusion of other control subject groups would be needed to confirm this.

The issue of practice is a large scale temporal relation. Temporal relations at smaller intervals also need to be discussed. The crucial aspect of timing is sometimes given surprisingly short shrift in studies of serial recall. It may seem silly to state the obvious, but the difference between presenting one item per every half second, and one item every two seconds is potentially huge given that estimates of the amount of information the phonological loop is able to hold reliably without active rehearsal is argued to be only about two seconds' worth of spoken material (Baddeley, 1986). Important is not only the delay period between the end of list presentation and the recall cue, but also the rate of item presentation, the duration of inter-item silence, as well as the delay exhibited between each element of the subject's response sequence. Use of computer presentation partially addresses these crucial timing issues. Note that although presentation timing was reliably monitored in this study, subject response was relatively self-paced. Although we do not report response times for subject reproductions here, this methodology conveniently allows for such measures to be gathered and analyzed. Somewhat more 'implicit' memory effects of the stimuli used might become evident if these measures were examined.

Note that even though response initiations were "relatively self-paced," the *duration* of the item response was constrained in that each stimulus sound needed to be played (and heard) in its entirety before the next response in the sequence could be initiated. This point illustrates another difference between this version of a reproduction task and one in which a subject is allowed to articulate at a self-selected rate. Having a respondent actively engage in articulating the reproduction, as opposed to causing a computer to generate the reproduction indirectly through using a set of subject-initiated button presses, are subtly different tasks. We would note that it has proven advantageous to accurately assessing phonological memory in certain clinical populations to compare multiple response formats for a particular span task. For example, clinical neuro-psychologists interpret greatly different spans between the traditional spoken response and a pointed response as suggesting a domain specific impairment (Lezak, 1995). We would simply reiterate the recommendation of using multiple response formats when testing memory capacity in the population of pediatric cochlear implant users.

One last point needs to be discussed regarding this test format. A closed-set choice of responses may seem very limiting in a number of ways, particularly in what types of error analysis can be attempted on subject responses. This is to be admitted, but there are advantages to using such as format as well. Using a closed set test requiring a set of motor responses makes the task a particularly easy one to administer, even with verbally shy or self-conscious children and reduces the risk of having the test be about the experimenter's perception of the response.

It is a cliché to state it, but clearly, the single most striking observation about short-term recall is its limits. From where and why does this limit originate? Are these limits primarily the result of activation interference? Trace decay? Or are they best understood as a byproduct of some limiting physiological process or some type of size/information load limitation in the neural machinery? What, if any, are the evolutionary advantages of its limitations? To what degree do these limits transcend sensory mode? Would it be more fruitful to attack these questions by talking of "acoustic attention" or even attention in the nonspecific sense, rather than phonological working memory? What models of neural activation best fit the sensory data on memory for acoustic stimuli? We discuss some of the development of the effects associated with immediate recall; what about their apparent breakdown during the aging process? Given that very few real world situations call for the demands of the task used here, that is, rapid recall of many similar lists ungoverned by any obvious higher structure, to what degree should we be concerned about the effects of proactive interference being responsible for the results? Does the fact that human memory system is not optimally designed to handle such situations reflect "the flip-side" of some other ability we use often and proficiently? Alan Baddeley has written that certain scientists "have accepted that digit span is not a good measure of the capacity of working memory, but have been less concerned about why this should be the case" (Baddeley, 1996). There is a good point buried in this: why should one type of information chunk necessarily be better than another type of chunk at estimating the general ability to store those chunks? That is, what kinds of information at what kind of time scales are we best and worst adapted to cope with and why might this be the case?

There is a chapter by McCarthy and Warrington (1990) whose title basically cuts to the heart of the matter: "Auditory-verbal span of apprehension: A phenomenon in search of a function." Are we actually studying an important human behavior function when we use simple memory span tasks, in light of the fact there have been individuals reported with very severely impaired simple spans who are able to converse and comprehend language, find their way around, and are generally able to function normally in their everyday lives? The common counter-argument would be that in selectively damaged brains individual compensatory strategies may be at work which tell us little about how the normal brain functions and develops. One might also point to the data that has been advanced by Gathercole and Adams (1993) that suggests that children that are able to efficiently encode and retrieve syllable strings, such as those used in novel words, are subsequently found to have faster growing vocabularies than their peers with shorter non-word repetition spans.

I am interested in how humans, particularly children, build the long-term representations of sound patterns that are used in word recognition. Recently, I have become interested in the ability to mimic speech sounds and its relation in the larger scheme of acquiring a spoken word and its meaning. Tasks such as non-word repetition appear to offer some insight into this process, from what I would argue is a "usefully narrow" perspective. Somewhat more temporally elongated reproduction tasks such as the one discussed in this paper should also tap into some of the same resources.

There are, relatedly, some interesting "trade-off" situations in early spoken word recognition. For example, words of short temporal duration place less of a load on limited memory storage resources, however, the longer the word, the more distinctive its phonological form and the easier it becomes to identify. In the case of cochlear implant users, for example, it has been demonstrated that long familiar multisyllabic words are easier for pediatric cochlear implant users to recognize than familiar short monosyllabic words (Kirk, Pisoni & Osberger, 1995). In other terms, the "difficulty" of holding early syllables in memory for purposes of confirming identification is overwhelmingly offset by the phonological distinctiveness of a multisyllabic word that can be confused with few other similar sounding words. When a spoken word is of low familiarity or entirely novel, however, this relationship is altered and perception on route to repetition/reproduction can be reasoned to require freshly encoding the heard stimulus in some form of a phonological store where the trace is weak and susceptible to "break-down" processes.

As hinted at in the introduction, there is currently some speculation that some of the unexplained outcome variability in open set spoken word recognition performance in prelingually-deafened pediatric cochlear implant users may share origins analogous to those thought to contribute to the distribution of good and poor readers in the population—namely, individual differences in the mechanisms of phonological immediate/working memory (Pisoni et al., 1997).

What data would one need to convincingly link the variability in cochlear implant use outcome and *modality non-specific* working memory capacity? A first step would be to show that of two groups of children, matched as closely as possible prior to implant except on performance on a battery of non-auditory working memory span tasks, (utilizing, for example, shape, color, or action responses), the group with more efficient "central executive" function made better progress on open set speech perception after implantation.

What data would one need to convincingly link outcome measures with a *speech modality-specific but not necessarily acoustic* memory capacity? In what ways can one probe for articulation-related memory ability in the profoundly-deaf pre-implant child? Does the extensive literature suggesting language-related learning in even the fetus and infant warrant probing for evidence of phonological and articulatory-visual memory mechanisms developed prior to onset of deafness? A review of the available research suggests that even prior to implant, conceivably because of "prelingual" learning experiences, many children utilize to some degree, aural-oral-related strategies to code and remember speech they are visually exposed to, while also using any manual/formational articulation strategies they may have acquired. The children that have managed to do this most successfully may be the ones that make the smoothest transition into using their implant for understanding speech, even without visual cues. Mogford (1987) suggests that while there are a number of studies attesting to low correlations between lip-reading ability and IQ, there is a relative dearth of data on how lip-reading behavior develops in very young children, both normal hearing and hearing-impaired. Some type of discrimination/habituation attention task involving visual presentation of speakers' faces may be what is required. I would favor this type of approach since interpreting communicative report from such young children seems fraught with hazards.

Many researchers and clinicians involved in cochlear implant research appear to believe that the currently observed differences in outcome can be accounted for in terms of hidden variables involving device hardware and implantation. When these improve, fewer children will fail to benefit from their implant. This is probably a safe bet. However, what is being downplayed in this version of the projected scenario is the possibility that the outcome distribution will simply shift; that is, on average, any given child will do better using the implant than in years past, however, there will still be children who do considerably better and considerably worse than the average pediatric implant user. Our hunch, stated again, is that outcome differences can be partially accounted for by the normal distribution of individual differences in the organization of those information processing circuits used in immediate phonological recall tasks. If this is indeed the case, it will have important implications for the type of individually designed post-implant rehabilitation programs recommended for individual children.

References

Allport, D. A. (1984). Auditory-verbal short-term memory and conduction aphasia. In H. Bouma & D. G. Bouwhuis (Ed.), *Attention and performance X: Control of language processes* (pp. 313-326). Hillsdale, NJ.

Baddeley, A. D. (1968). How does acoustic similarity influence short-term memory? *Quarterly Journal of Experimental Psychology*, **20**, 249-264.

Baddeley, A. D. (1986). *Working-memory*. London: Oxford University Press.

Baddeley, A. D. (1992). Working memory. *Science*. **255**, 556-559.

Baddeley, A. D. (1990). The development of the concept of auditory-verbal short-term memory: Implications and contributions of neuropsychology. In G. Vallar & T. Shallice (Eds.), *Neuropsychological impairments of short-term memory*. New York: Cambridge University Press.

Baddeley, A. D. (1996). The concept of working memory. In S. E. Gathercole (Ed.), *Models of short-term memory* (pp.1-27). UK: Psychology Press LEA.

Baddeley, A. D. & Hitch, G. (1974). Working memory. In G. H. Bower (Ed.), *The psychology of learning and motivation, Volume 8*. New York: Academic Press.

Baddeley, A. D., Thomson, N. & Buchanan, M. (1975). Word length and the structure of short-term memory. *Journal of Verbal Learning and Behavior*, **14**, 575-589.

Bess, F.H. & Humes, L. E. (1990). *Audiology: The fundamentals*. Baltimore: Williams & Wilkins.

Best, C. T. (1994). The emergence of native-language phonological influences in infants: A perceptual assimilation model. In Goodman, J. C., & Nusbaum, H. C. (Eds.), *The development of speech perception: The transition from speech sounds to spoken words* (pp. 167-224). Cambridge, MA: MIT Press.

Bower, G. H., Clark, M. C., Lesgold, A. M., & Winzenz, D. (1969). Hierarchical retrieval schemes in recall of categorized word lists. *Journal of Verbal Learning and Verbal Behavior*, **8**, 323-343.

Campbell, R. (1987). Lip-reading and immediate memory processes or on thinking impure thoughts. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lip-reading* (pp.243-255). London: Lawrence Erlbaum.

Campbell, R. & Wright, H. (1990). Deafness and immediate memory for pictures: Dissociation between inner speech and inner ear. *Journal of Experimental Child Psychology*, **50**, 259-286.

Case, R. D., Kurland, D.M., & Goldberg, J. (1982). Operational efficiency and the growth of short-term memory span. *Journal of Experimental Child Psychology*, **33**, 386-404.

Cole, R. A. (1973). Different memory functions for consonants and vowels. *Cognitive Psychology*, **4**, 39-54.

Conrad, R. (1964). Acoustic confusions in immediate memory. *British Journal of Psychology*, **55**, 75-84.

Conrad, R. (1970). Short-term memory processes in the deaf. *British Journal of Psychology*, **61**, 179-195.

Conrad, R. (1972). Short-term memory in the deaf: A test for speech coding. *British Journal of Psychology*, **63**, 173-180.

Conrad, R. & Hull, A. J. (1964). Information, acoustic confusion and memory span. *British Journal of Psychology*, **55**, 429-432.

Conway, A.R.A. & Engle, R.W. (1996). Individual differences in working memory capacity: More evidence for a general capacity theory. *Memory*, **4**, 577-590.

Corsi, P. M. (1972). Human memory and the medial temporal region of the brain. Unpublished doctoral dissertation, McGill University.

Crowder, R.C. & Surprenant, A. M. (1995). On the linguistic module in auditory memory. In De Gelder, B. & Morais, J. (Eds.), *Speech and reading: A comparative approach* (pp. 49-64). Hove, England: Erlbaum (UK)/Taylor & Francis.

Dallett, K. (1964). Intelligibility and short-term memory in the repetition of digit strings. *Journal of Speech and Hearing Research*, **7**, 362-368.

Daneman, M. & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, **19**, 450-466.

Daneman, M., Nemeth, S., Stainton, M., & Huelsmann, K. (1995). Working memory as a predictor of reading achievement in orally educated hearing-impaired children. *The Volta Review*, **97**, 225-241.

Daneman, M. & Tardif, T. (1987). Working memory and reading skill re-examined. In M. Coltheart (Ed.), *Attention and performance XII* (pp.2491-508). London: Erlbaum.

Darwin, D.J. & Baddeley, A.D. (1974). Acoustic memory and the perception of speech. *Cognitive Psychology*, **6**, 41-60.

Dedina, M.J. (1987). SAP: A Speech acquisition program for the SRL-VAX. In *Research on Speech Perception Progress Report No. 13*. Bloomington IN: Speech Research Laboratory, Indiana University, 331-337.

Dodd, B., Hobson, P., Brasher, J. & Campbell, R. (1983). Short-term memory in deaf children. *British Journal of Developmental Psychology*, **1**, 353-364.

Dollaghan, C. & Campbell, T. F. (1997). Nonword repetition and child language impairment. Unpublished manuscript.

Drewnowski, A. (1980). Memory functions for vowels and consonants: A reinterpretation of acoustic similarity effects. *Journal of Verbal Learning and Behavior*, **19**, 176-193.

Drewnowski, A. & Murdock, B. B. (1980). The role of auditory features in memory span for words. *Journal of Experimental Psychology: Human Learning and Memory*, **6**, 319-332.

Ellis, N. C. & Hennelly, R. A. (1980). A bilingual word-length effect: Implications for intelligence testing and the relative ease of mental calculation in Welsh and English. *British Journal of Psychology*, **71**, 43-51.

Ericsson, K. A. & Lehmann, A. C. (1996). Expert and exceptional performance: Evidence of maximal adaptation to task constraints. *Annual Review of Psychology*, **47**, 273-305.

Gathercole, S. (1987). Lip-reading: Implications for theories of short-term memory. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lip-reading* (pp.227-241). London: Lawrence Erlbaum.

Gathercole, S. & Adams, A. (1993). Phonological working memory in very young children. *Developmental Psychology*, **29**, 770-778.

Gathercole, S. E., Adams, A., & Hitch, G. J. (1994). Do young children rehearse? An individual-differences analysis. *Memory and Cognition*, **22**, 201-207.

Gathercole, S.E. & McCarthy, R. A. (1994). *Memory tests and techniques. Special issue of the journal Memory*, **2**, Hillsdale, NJ: Lawrence Erlbaum Associates.

Goldinger, S. D., Pisoni, D. B. & Logan, J. S. (1991). On the nature of talker variability effects on recall of spoken word lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **17**, 152-162.

Goldman-Rakic, P. S. (1996). Regional and cellular fractionation of working memory. *Proceedings of the National Academy of Sciences of the United States of America*, **93**, 13473-13480.

Haber, R. N. (1969). *Information-processing approaches to visual perception*. Holt, New York: Rinehart & Winston.

Hanson, V. L. (1982). Short-term recall by deaf signers of American Sign Language: Implications for order recall. *Journal of Experimental Psychology: Learning, Memory and Cognition*, **8**, 572-583.

Hanson, V. L. & Lichenstein, E. H. (1990). Short-term memory coding by deaf signers: The primary language coding hypothesis reconsidered. *Cognitive Psychology*, **22**, 211-224.

Hernández, L.R. (1994). Implementation of a PC-based perceptual testing system (PTS): A first milestone. In *Research on Spoken Language Processing Progress Report No.19.* Bloomington, IN: Speech Research Laboratory, Indiana University, 321-328.

Hernández, L.R. (1995). Current computer facilities in the Speech Research Laboratory. In *Research on Spoken Language Processing Progress Report No.20.* Bloomington, IN: Speech Research Laboratory, Indiana University, 389-393.

Hirsh-Pasek, K. & Treiman, R. (1982). Recoding in silent reading: Can the deaf child translate print into a more manageable form? *The Volta Review*, **84**, 71-82.

Hitch, G. J. & Halliday, M. A. (1983). Working memory in children. *Philosophical Transactions of the Royal Society: London*, **B302**, 325-340.

Hood, J.D. & Poole, J.P. (1980). Influence of the speaker and other factors affecting speech intelligibility. *Audiology*, **19**, 434-455.

Hulme, C. (1984). Developmental differences in the effects of acoustic similarity on memory span. *Developmental Psychology*, **20**, 650-652.

Hulme, C., Maughan, S., & Brown, G. D. (1991). Memory for familiar and unfamiliar words: Evidence for a long-term memory contribution to short-term memory span. *Journal of Memory and Language*, **30**, 685-701.

Hulme, C. & Roodenrys, S. (1995). Verbal working memory development and its disorders. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, **36**, 373-398.

Hulme, D., Roddenrys, S., Brown G., & Mercer, R. (1995). The role of long term memory mechanisms in memory span. *British Journal of Psychology*, **86**, 527-536.

Humes, L. E. & Christopherson, L. (1991). Speech identification difficulties of hearing-impaired elderly persons: The contributions of auditory processing deficits [Abstract]. *Journal of Speech and Hearing Research*, **34**, 686-693.

Humphreys, M.S., Lynch, M. J., Revelle, W., & Hall, J.W. (1983). Individual differences in short-term memory. In R. F. Dillon & R. R. Schmeck (Eds.), *Individual differences in cognition: Volume I* (pp. 35-64). New York: Academic Press.

Jonides, J., Reuter-Lorenz, P. A., Smith, E., Awh, E., Barnes, L. L., Drain, M., Glass, J., Lauber, E. J., Patalano, A. L., & Schumacher, E. H. (1996). Verbal and spatial working memory in humans. In *The psychology of learning and motivation, Volume 35*, San Diego: Academic Press.

Kirk, K.I., Pisoni, D.B. & Osberger, M.J. (1995). Lexical effects on spoken word recognition by pediatric cochlear implant users. *Ear and Hearing*, **16**, 470-481.

Levitt, H. (1971). Transformed up-down methods in psychoacoustics. *The Journal of the Acoustical Society of America*, **49**, 467-477.

Lezak, M. D. (1995*). Neuropsychological assessment.* New York: Oxford University Press.

Logan, J. S. (1992). *A Computational Analysis of Young Children's Lexicons*. Research on Spoken Language Processing, Technical Report No. 8, Bloomington, IN: Speech Research Laboratory, Indiana University.

Logie, R. H., Della Salla, S., Laiacona, M., Chalmers, P., & Wynn, V. (1996). Group aggregates and individual reliability: The case of verbal short-term memory. *Memory and Cognition*, **24**, 305-321.

Luce, P. A. (1986). *Neighborhoods of Words in the Mental Lexicon*. Research on Speech Perception Technical Report No. 8, Bloomington, IN: Speech Research Laboratory, Indiana University.

Luce, P.A. & Pisoni, D.B. (1998). Recognizing spoken words: the Neighborhood Activation Model. *Ear and Hearing*. **19**, 1-39.

Luce, P. A., Feutstel, T. C. & Pisoni, D. B. (1983). Capacity demands in short-term memory for synthetic and natural speech. *Human Factors*, **25**, 17-32.

Marschark, M. & Mayer, T. S. (In press, 1997). Mental representation and memory in deaf adults and children. In M. Marschark & M. D. Clark (Eds.), *Psychological perspectives on deafness: Volume 2*. Mahway, NJ: Lawrence Erlbaum.

Martin, C.S., Mullennix, J.W., Pisoni, D.B., & Summers, W.V. (1989). Effects of talker variability on recall of spoken word lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **15**, 676-684.

Martin, R. C. & Breedin, S. D. (1992). Dissociations between speech perception and phonological short-term memory deficits. *Cognitive Neuropsychology*, **9**, 509-534.

McCarthy, R. A. & Warrington, E. K. (1990). Chapter 7: Auditory-verbal span of apprehension: a phenomenon in search of a function. In G. Vallar & T. Shallice, (Eds.), *Neuropsychological impairments of short-term memory*. Cambridge: Cambridge University Press.

Mendel, L. L. & Danhauer, J. L. (1997). *Audiologic evaluation and management and speech perception assessment*. San Diego: Singular.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits to our capacity for processing information. *Psychological Review*, **63**, 81-97.

Miller, G. & Nicely, P. E. (1955). An analysis of perceptual confusions among some English consonants. *Journal of the Acoustical Society of America*, **27**, 338-352.

Milner, B. (1971). Interhemispheric differences in the localization of psychological processes in man. *British Medical Bulletin*, **27**, 272-277.

Mogford, K. (1987). Lip-reading in the prelingually deaf. In B. Dodd & R. Campbell (Eds.) *Hearing by eye: The psychology of lip-reading* (pp. 191-211). London: Lawrence Erlbaum.

Naveh-Benjamin, M. & Ayres, T. J. (1986). Digit span, reading rate, and linguistic relativity. *Quarterly Journal of Experimental Psychology*, **38A**, 739-751.

Neisser, U. (1967). *Cognitive psychology*. New York: Appleton-Century-Crofts.

Papagno, C. & Vallar, G. (1992). Phonological short-term memory and the learning of novel words: The effect of phonological similarity and item length. *The Quarterly Journal of Experimental-Psychology: Human Experimental Psychology*, **44A**, 47-67.

Perfetti, C. A. & Lesgold, A. M. (1977). Discourse comprehension and sources of individual differences. In M. A. Just & P. A. Carpenter (Eds.), *Cognitive processes in comprehension*. Hillsdale, NJ: Lawrence Erlbaum.

Pisoni, D. B. (1975). Auditory short-term memory and vowel perception. *Memory and Cognition*, **3**, 7-18.

Pisoni, D. B., Svirsky, M. A., Kirk, K. I. & Miyamoto, R. T. (1997). Looking at the "Stars": A first report on the intercorrelations among measures of speech perception, intelligibility, and language in pediatric cochlear implant users. Draft of a paper presented at the Vth International Cochlear Implant Conference, May 1-3, 1997, New York, NY.

Rabbitt, P. M. A. (1968). Channel-capacity, intelligibility, and immediate memory. *Quarterly Journal of Experimental Psychology*, **20**, 241-248.

Rayner, K. & Pollatsek, A. (1989). *The psychology of reading*. Hillsdale, NJ: Lawrence Erlbaum.

Roodenrys, S., Hulme, C., Alban, J., Ellis, A. W. & Brown, G. (1994). *Memory and Cognition*, **22**, 695-701.

Schweickert, R. (1993). A multinomial processing tree model for degradation and redintegration in immediate recall. *Memory and Cognition*, **21**, 168-175.

Shand, (1982). Sign-based short-term coding of American Sign Language signs and printed English words by congenitally deaf signers. *Cognitive Psychology*, **14**, 1-12.

Snowling, M. J. (1981). Phonemic deficits in developmental dyslexia. *Psychological Research*, **43**, 219-234.

Sumby, W. H. (1963). Word frequency and serial position effects. *Journal of Verbal Learning and Verbal Behavior*, **1**, 443-450.

Swanson, H. L. (1996). Individual and age-related differences in children's working memory. *Memory and Cognition*, **24**, 70-82.

Tehan, G. & Humphreys, M. S. (1988). Articulatory loop explanations of memory span and pronunciation rate correspondences: A cautionary note. *Bulletin of the Psychonomic Society*, **26**, 293-296.

Treiman, R. & Danis, C. (1988). Short-term memory errors for spoken syllables are affected by the linguistic structure of the syllables. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **14**, 145-152.

Turner, M. L. & Engle, R. W. (1989). Is working memory task dependent? *Journal of Memory and Language*, **28**, 127-154.

Vallar, G. & Papagno, C. (1986). Phonological short-term store and the nature of the recency effect: Evidence from neuropsychology. *Brain and Cognition*, **5**, 428-442.

Vallar, G. & Shallice, T. (1990). *Neuropsychological impairments of short-term memory.* New York, NY: Cambridge University Press.

Watson, C. S. (1991). Auditory perceptual learning and the cochlear implant. *American Journal of Otology*, **12 Suppl**, 73-79.

Wechsler, D. (1955). *Wechsler adult intelligence scale.* New York: The Psychological Corporation.

Wechsler, D. (1991). *Wechsler intelligence scale for children—Third edition.* San Antonio: The Psychological Corporation.

Wetherick, N. E. (1975). The role of semantic information in short-term memory. *Journal of Verbal Learning and Verbal Behavior*, **14**, 471-480.

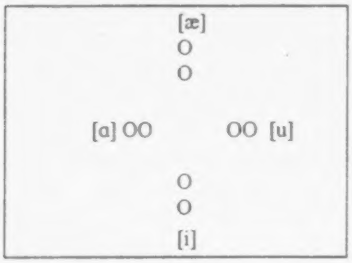
Wickelgren, W.A. (1965). Short-term memory for phonemically similar lists. *American Journal of Psychology*, **78**, 567-574.

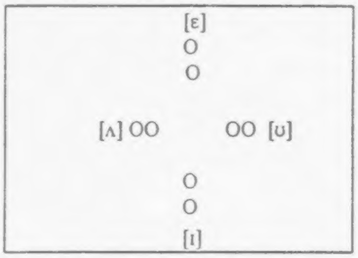
Woodcock, R.W. & Johnson, M. B. (1977). *Psychoeducational battery. Part two: Tests of achievement.* Hingham, MA: Teaching Resources.

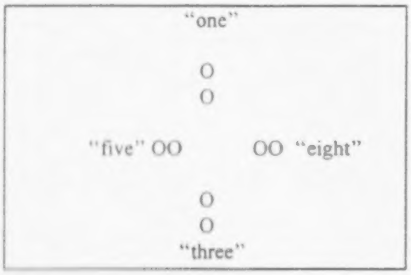
Appendix A

Mapping Between Stimuli and Button Locations

(Inner diamond corners are lights; outer diamond corners are buttons.)


[æ]
O
O
[ɑ] OO
OO [u]
O
O
[i]


[ɛ]
O
O
[ʌ] OO
OO [ʊ]
O
O
[ɪ]


"one"
O
O
"five" OO
OO "eight"
O
O
"three"

Appendix B

Instructions

Welcome to the Speech Research Laboratory. The experiment you will be participating in today is part of a larger project concerned with short-term and working memory for auditory sequences.

The stimuli you will hear through your headphones are sounds recorded from a male speaker of American English. The first two sections of this experiment use vowel-sound stimuli. The last section uses spoken digit-name stimuli.

There will be three main sections to this experiment. Within each of the three sections, there will be four "stages". Before each stage you will receive a reminder on-screen that will tell you which stage is next.

During the first stage, **"Familiarize"**, you will hear a sequence of sounds through your headphones accompanied by a matching light sequence on a button response box. There are four different sounds, and each sound will be played twice. **During "Familiarize", your job is just to watch and listen. Try to remember which sounds go with which light/button.**

During each trial in the second stage, **"Learn"**, you will hear a single sound through your headphones. **Your job is to firmly press the button that was matched with that particular sound during familiarization.** If you press the correct button, the light will go on, and the matching sound will play. If you press an incorrect button, no sound at all will play, and a light at the *correct* location will flash on. This feedback is designed to help you learn the correct mapping of sounds to buttons.

The third stage is called **"Practice"**. During each trial in this stage, you will again hear a single sound through your headphones. **Your job again is to press the button that matches the sound.** If you press the correct button, the light will go on and the matching sound will play. However, during this stage, if you press an incorrect button, the sound matched to that button will play and its light will flash on. You will receive no feedback about which was the correct response.

The fourth and last stage is called **"Sequence"**. During each trial in this stage, you will hear a sequence of one or more sounds through your headphones. Some stimuli sequences will be short, others will be longer. **Your job is to replicate the sequence of sounds by pressing the appropriate buttons.** Again, remember to press firmly on each button. By waiting after you have finished entering your responses, you will advance to the next trial. If you realize that you have made a mistake in your selection of presses, simply ignore the mistake and finish entering the remainder of your response (with the correction, if you wish). Wait to advance to the next trial.

After each section containing all four of these stages, an "Intermission" message will appear on the screen. Please sit quietly, and when you feel ready to do so, press any button to resume the experiment. Please work at a quick but comfortable pace. Feel free to consult these directions at any break in the experiment. When the screen display indicates that the experiment is over, please remove your headphones and remain seated. The experimenter will indicate when the session is officially finished.

Please ask any questions concerning the above material now. Then put this sheet aside to indicate that you have finished reading.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 21 (1996-1997)
Indiana University

Static vs. Dynamic Faces as Retrieval Cues in Recognition of Spoken Words[1]

Lorin Lachs

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405

---

[1] This research was supported by NIH-NIDCD Research Grant DC00111 and NIH-NIDCD Training Grant DC00012 to Indiana University.

Static vs. Dynamic Faces as Retrieval Cues in Recognition of Spoken Words

Abstract. Three experiments examined the integration of auditory and visual information in memory for spoken words. Across experiments, recognition of isolated words was tested in the context of studied or non-studied faces and voices. The degree to which faces were informative about the studied speech event was manipulated between experiments. In Experiment 1, faces were static pictures experimentally paired to the voice speaking the word. In Experiment 2, a control for Experiment 1, faces were presented upside down. In Experiment 3, faces were dynamic video clips of talkers articulating. Subjects were either instructed that faces were to be recognized explicitly or not. The results show that static faces can only be used as effective retrieval cues to the recognition of words when experimental conditions encourage the association between visual and auditory information. By contrast, dynamic, articulating faces are automatically encoded along with voice and word information and improve recognition performance substantially. In addition, the encoding of faces for explicit recognition interferes with the utility of dynamic faces in recognizing speech. The results are taken to imply that visual information is encoded in cross-modally integrated memory representations for speech, but only when it is informative about the speech event.

Introduction

Although much of the research on speech perception and spoken word recognition conducted in the past has regarded speech as a purely auditory phenomenon, a growing body of evidence suggests that the visual sensory modality plays an important role in the perception of speech and the understanding of spoken language. Perhaps the most well-known phenomenon relevant to this area of inquiry is the so-called "McGurk effect", discovered by McGurk and MacDonald (1976). They found that the perception of speech can be altered by simultaneously presenting conflicting information in the auditory and visual modalities. Specifically, they reported that the visual presentation of a talker's face repeatedly articulating the syllable [ga], along with the dubbed audio presentation of the same talker speaking the syllable [ba], induced in subjects the perception of the syllable [da] (McGurk & MacDonald, 1976). This effect is very robust; 98% of adults who viewed stimuli of this nature reported a "fused" percept (i.e., the perception of the syllable [da] given an auditory [ba] and a visual [ga]). Furthermore, McGurk and MacDonald (1976) reported that the effect did not habituate over time; despite long experience with and full knowledge of the nature of the stimuli, the authors themselves continued to experience the effect.

Since its discovery, a great deal of research has been conducted on the nature of the McGurk effect. In study after study, the effect has been replicated, using various experimental manipulations (MacDonald & McGurk, 1978; Massaro & Cohen, 1983; Massaro & Cohen, 1990; Rosenblum & Saldaña, 1992; Summerfield & McGrath, 1984). For instance, Summerfield and McGrath (1984) found that the effect is not necessarily confined to the perception of consonants. By presenting simultaneous conflicting information in the auditory and visual modalities, Summerfield and McGrath (1984) reported that subjects perceived vowels which were combinations of the vowels specified by either modality alone. Additionally, Dekle, Fowler & Funnel (1992) found that audiovisual integration can occur in the perception of real words. In another study, Massaro (1987) found that explicitly instructing subjects to ignore one of the sensory dimensions does not eliminate the effect it has on the perception of the other dimension. Indeed, the research into this phenomenon has been extensive. Among the copious amount of research into the

McGurk effect is research which shows, for example, that the asynchronous presentation of the information in the two modalities still elicits an effect (Munhall, Gribble, Sacco, & Ward, 1996) and that, for certain syllables, the effect can be evoked with inverted articulating faces (Campbell, 1994; Massaro & Cohen, 1996).

Taken together, the research findings on the McGurk effect have produced compelling evidence that the influence of the visual sensory modality on speech perception is substantial and worthy of study. Indeed, Summerfield (1987) lists the McGurk effect as one of five major phenomena that must be accounted for by any theory of speech perception. Still, doubts as to the significance of the McGurk effect may be raised, due to its extremely artificial nature. In the real world, one is hardly, if ever, confronted with a situation in which one must perceive speech in the presence of conflicting, non-degraded information from both the auditory and visual modalities. Fortunately, more naturalistic evidence of vision's role in the process of speech perception comes from a groundbreaking study which pre-dates McGurk and McDonald (1976) by over twenty years and provides the foundation upon which all audiovisual theories of speech perception may stand.

In order to assess the contribution of visual information to the process of speech perception, Sumby and Pollack (1954) had two groups of subjects seated around a talker. Each listener wore a pair of headphones and sat rather close to the talker. Half of the listeners, however, could see the talker, while the other half, turned away, could not. The words spoken during all trials were mixed with white noise at varying signal-to-noise ratios. After each stimulus, subjects made a forced-choice response from a list of words whose length varied between subjects.

Sumby and Pollack (1954) observed several important relations. First, they found that the intelligibility of words spoken in noise is negatively affected by the size of the possible message set. This replicated the well-known result that as the number of possible words increases, the susceptibility of those words to interference by noise increases (Miller, Heise, & Lichten, 1951).

Second, Sumby and Pollack found that the average intelligibility of *audiovisually* presented words from different vocabulary sizes, when plotted as a function of speech-to-noise ratio, were substantially higher than the unimodal scores. For example, under the most degraded condition tested (-30 dB S/N), the average intelligibility of the eight word list went from about 15% correct in the unimodal condition to around 90% correct in the multimodal condition. Put in other terms, the size of the gain in speech intelligibility as a result of multimodal presentation was roughly equal to the gain in intelligibility afforded by an increase in signal-to-noise ratio of +15 dB (Erber, 1969; MacLeod & Summerfield, 1987; Middleweerd & Plomp, 1987; Rosenblum & Saldaña, 1996; Sumby & Pollack, 1954).

Interestingly, this remarkable gain in accuracy as a result of multisensory presentation was found to interact with the size of the possible message set. Greater advantages were found for small vocabulary sizes under very degraded conditions. Furthermore, this interaction was not simply a result of the fact that, probabilistically, smaller set sizes lead to higher accuracies, since, when the results were plotted in terms of the percent information gained (a measure which normalizes for stimulus set size), the effect was still observed. Sumby and Pollack found, however, that for higher signal-to-noise ratios the visual contribution to intelligibility was mainly detected with longer vocabulary lists, whereas little if any advantage was gained for smaller vocabulary lists. Although this may seem puzzling at first, the results (as Sumby and Pollack (1954) point out) appear to be due mainly to the fact that at the higher signal-to-noise ratios (i.e., under less degraded conditions) unimodal performance for small vocabulary sizes is already near ceiling levels, thereby leaving little room for improvement as a result of additional visual information.

In light of this "ceiling effect," Sumby and Pollack (1954) proposed that the measure of visual information's contribution to speech intelligibility should be scaled in terms of its *possible* contribution. Remarkably, when they re-analyzed the data, they found that this measure, which they called "R" (i.e., the ratio of the actual contribution of visual information to its possible contribution), remained constant across a wide range of signal-to-noise ratios. In other words, the relative contribution of visual information was found to be independent of the signal-to-noise ratio under test. This finding can be interpreted to mean that the utility of the information provided by vision is not simply "additional" to that provided by audition, but is instead somehow intrinsic to the process of speech perception itself: degrading the auditory signal with noise only serves to tease apart the underlying contributions of both input modalities to the perceptual process.

In another intelligibility experiment, Erber (1969) replicated the initial results of Sumby and Pollack (1954) and extended them in several ways. Erber (1969) reported that while the threshold for performance above chance in auditory-only conditions lies somewhere around -18 dB S/N, performance in audiovisual conditions begins to increase above baseline lip-reading at a much lower signal-to-noise ratio, implying that visual information can work in tandem with auditory information in the perception of speech, even when the information in the auditory signal alone would be unusable to the listener.

Taken together, these early studies have demonstrated that the information which is specified by the visual modality can have a major influence on speech perception and spoken word recognition. Subsequent studies have attempted to determine the exact nature of this influence. Many of these studies have concentrated on the fact that the aspects of acoustically transmitted speech which are most confusable in noise are precisely those aspects of the message that are reliably transmitted visually (Massaro, 1987; Summerfield, 1987). For example, while the phonemes /k/ and /p/ are confusable at signal-to-noise ratios of +12 dB and below when presented acoustically (i.e., at a very low level of noise), cluster analysis of confusions made by subjects when identifying visual only speech reveals that these particular visual phonemes ("visemes") are among the last consonants to be confused when presented visually (Summerfield, 1987). The exact process by which these stimulus properties are exploited and integrated is a topic of some debate; Summerfield (1987) takes these properties as support for the assertion that information from both modalities is integrated before categorization, while Massaro (1987) claims that these confusions point to evidence that sub-phonetic categorical judgments are made on modality specific inputs which are then integrated *after* perceptual analysis.

Since both of these possibilities seem equally likely from the viewpoint of perceptual studies, it has become necessary to examine the nature of multimodal speech representations in memory, since insight into the way in which information from disparate modalities is encoded may provide important new clues about the underlying process by which this information is obtained from sensory input.

In a recent study of the effects of multimodal input on memory, Pisoni, Saldaña, and Sheffert (1995) showed an influence of audiovisual encoding on two aspects of memory: immediate memory span and serial recall. In the immediate memory span experiment, Pisoni et al. (1995) found that the number of items which could be recalled correctly was significantly shorter when stimulus items were presented audio-visually, compared with audio-only presentations. This result was taken to imply that visual information usurps processing resources available to the limited capacity working memory system. In the serial memory experiment, Pisoni et al. found that items in the primacy portion of the list were recalled more accurately when they were originally presented audio-visually than when they were originally presented audio-only. Since performance on items in the primacy portion of serial recall lists is usually taken to reflect the degree to which those items have been rehearsed, encoded and transferred to long term memory, Pisoni et al. concluded that the additional visual information must also be encoded into long term memory along with the auditory information and can be used as an effective cue at the time of retrieval.

Taken together, these recent findings suggest that visual information about a speech event is processed and encoded in some way relating to the phonetic information provided by the same event. These findings do not, however, provide an answer to whether the information from the two modalities is stored in a multimodal, integrated form or in separate, but linked, unimodal representations appropriate to each input modality.

One study designed to provide insight into this question was conducted recently by Sheffert and Fowler (1995). Using a continuous recognition task, Sheffert and Fowler (1995) examined whether words could be more accurately recognized at test when presented along with the studied video information about the talker. Sheffert and Fowler found that, while presentation of the word at test using the same voice always facilitated recognition of the word, little or no advantage was gained as a result of repeated video contexts. Several other studies were conducted to assess the degree to which visual information was encoded but all of these revealed the same basic pattern of results: repeated voices facilitated the recognition of words, while repeated faces did not. Sheffert and Fowler therefore concluded that voices have a "privileged" status in the mnemonic encoding of words, and that faces may play a more contextual role.

In another effort to examine the question of multimodal integration in memory, Kato, Kanzaki, Tohkura, and Akamatsu (1995) examined the recognition of spoken sentences presented in one modality given changes in the stimulus presented in the other modality. Their study was motivated by the earlier findings of Legge, Grosmann, and Pieper (1984), who showed that presenting a static picture of a face during the study interval for a particular voice increased the probability that that voice would be recognized later. However, the Legge et al. (1984) study did not examine whether voices could aid in the recognition of faces, a question of critical importance to the debate over integrated multimodal encoding. If it were found that faces can aid in the recognition of voices (as demonstrated by Legge et al.) but not vice versa, then this would be strong evidence against the notion that information from the two modalities is stored in an integrated representation.

Kato et al. (1995) therefore designed a recognition memory experiment to examine this issue. During the study phase, subjects were presented with a static picture of a face and a concurrent recording of a voice speaking a sentence. Throughout the study phase, the specific face+voice pairings remained constant, although the pairing was assigned randomly between subjects (i.e., Face A was not necessarily presented along with Voice A for all subjects); each face+voice pairing was seen six times (two sentences were presented three times). During the test phase, subjects were asked whether the face (or voice) that was presented was a face that they had already seen (or, alternatively, whether the voice was a voice which they had already heard). The item in the test modality (i.e., the face or the voice, depending on whether the task was to recognize the face or the voice, respectively) was presented in one of four "other-mode contexts." So, for example, if the test mode was the face, then an *old* (i.e., studied) face could appear with either the voice with which it had been presented at study, a different voice that had been paired with another face during study, a different voice which the subjects had not yet heard, or no concurrent voice information. Similarly, a *new* face could appear with either a studied voice, a new voice which they had never heard, or no concurrent voice information. This experimental design was employed to assess whether a face or voice could be used as a facilitatory cue for retrieval of the other mode.

BEST COPY AVAILABLE

Several measures of performance were obtained. First, the hit rate (the rate of correctly identifying the test mode as "old" when given a studied, "old" item) was examined. Kato et al. (1995) found that while there was no difference in accuracy between correctly recognizing voices and faces, there was a significant effect of the other-mode context on recognition accuracy. That is, subjects were equally able to recognize faces and voices, but this ability was affected by the simultaneously presented other-mode context, due to decreased performance when the other-mode was *not* studied with the test item, relative to when no other-mode context was available (Kato et al., 1995).

Consider the following example as an illustration of the above finding: say that during the study phase for a particular subject, Face A was always presented with Voice A and Face B was always presented with Voice B. If this subject's performance was consistent with the results of Kato et al. (1995), then his recognition accuracy for Face A would be worse in the context of Voice B than when there was no voice at all. However, the recognition of Face A in the context of Voice A or Voice B would not be different. In other words, recognition of a face was best when no voice context was given; if a voice context *was* given, then whether or not that voice context was the one with which the face had been originally studied did not significantly impact recognition performance. This pattern of results suggests that static faces and voices are not stored integrally, but instead that voices can interfere with the recognition of faces. In a second analysis, Kato et al. (1995) examined the correct rejection rate (the rate of correctly identifying the test mode as "new" when given a non-studied item). Here they found that the rejection of a new *face* was not affected by the other-mode context, while the rejection of a *voice* was facilitated in the presence of new faces (Kato et al., 1995).

Kato et al. (1995) concluded that the pattern of their results suggests that face and voice information are encoded independently in memory, since recognition of a test item was not better in the context of its correct other mode item, relative to performance in the context of incorrect or novel other mode items (i.e., there was no facilitation).

There are, however, several weaknesses in the methodology and design of this experiment that call into question their conclusion. First of all, the study used an explicit recognition memory procedure to assess the encoding of face and voice in memory. That is, subjects were asked explicitly on each presentation to determine whether or not the face or the voice was a component attribute of the stimulus item that was presented during the study phase. This task has no reference to the other mode item; in other words, there is no reason for the subject to encode both modalities in an integrated representation, since their only reference to each other is simply an arbitrary co-occurrence or association during the study phase.

Similarly, because the pairing of faces and voices was randomly assigned as a between subjects variable, there was no reason, other than the arbitrary experimental ones, that a subject would encode the face and voice together in an integrated representation in memory.

Finally, and perhaps most importantly, the use of static pictures of faces eliminates any naturally occurring dynamic optical information that links faces and voices together. That is, by eliminating dynamic, articulatory information from the optical display, all aspects of the relationship between the face and the voice speaking were eliminated from the stimulus, once again leaving only experimental conditions to signify the importance of the intermodal pairing.

The present series of experiments was motivated by the Kato et al. (1995) experiment and was designed to deal with the criticisms mentioned above by using dynamic visual displays and examining both implicit and explicit memory processes.

Experiment 1

In Experiment 1, the task was changed from an explicit recognition memory test of faces or voices to an implicit one; on each trial during the test phase, subjects were required to recognize whether the *word* they heard was an item that was presented earlier during the study phase. Faces and voices were manipulated but their effects were never expressed explicitly. It should also be noted that while sentences acted as the carrier for voice information in the Kato et al. (1995) study, the present study utilized isolated words as both the carrier of voice information *and* the test items used for recognition.

The test items could be presented in one of four different conditions. In the (F+V+) condition, the face and voice which were presented with the word during the study phase were also presented during the test phase. In the (F+v-) condition, the face was the same as during study, but the voice was not. In the (f-V+) condition, the voice was the same as during study, but the face was not. Finally, in the (f-v-) condition, the face and voice were both different from those presented with the word during study. The prediction was that these experimental manipulations would allow a more sensitive measure of the encoding of face and voice in memory; in addition, these manipulations allowed the effects of face and voice on recognition memory to be examined independently.

The main purpose of Experiment 1 was to serve as a control by replicating the findings of Kato et al. (1995) using a new implicit memory task and different stimuli. A replication of the major results was expected because, in this experiment, there was still no reason other than experimental necessity for a subject to associate a particular face with a particular voice: the faces were still static and semi-randomly assigned to one another at the time of study[2].

Because the results of Kato et al. (1995) are not directly interpretable within the implicit memory paradigm, we should be clear about what was expected. Using subjects' explicit judgments on face and voice recognition, Kato et al. (1995) found some evidence that intermodal relationships could exist in memory (e.g., better rejection of new voices in the presence of new faces, and inhibitory effects on recognition due to incongruent other-mode stimuli) and interpreted the absence of facilitatory effects to mean that these intermodal relationships existed as links between independent, unimodal memory representations. Our task, on the other hand, used subjects' recognition of *words* to examine their knowledge of face+voice pairings. Still, the change in task should not affect the basic finding reported in Kato et al. (1995) that intermodal relationships can indeed exist in memory. As such, we expected that Experiment 1, if it were a valid extension of the findings in Kato et al. (1995) should indeed show an effect of modality context on recognition of words.

Before work could begin on this experiment, it was important to make sure that subjects were actually encoding and using the visual information during study and test. It is possible that subjects might just as well have closed their eyes and performed the task on the basis of word and voice information alone.

---

[2] The assignment of faces to voices was semi-random due to the fact that, in our stimulus database, half of the talkers are male and half of them are female. Since all of the talkers in Kato et al. (1995) were male, throwing in the mixing of male voices with female faces and vice versa could be a potential confound in the replication of the original study. As such, only male voices were assigned to male faces, and only female voices were assigned to female faces. Within these limitations, however, the assignment of a particular face to a particular voice was randomized between subjects.

As a result, our measures might be rendered useless. In order to control this situation, two additional manipulations were introduced into the experimental design. First, at the end of the test phase, subjects were presented with a test of explicit face recognition. Performance on this additional task was used as a criterion for inclusion into the final data analysis. Second, we also manipulated the encoding instructions. For half of the subjects, advance notice of the explicit face recognition test at the end of the session was given directly in the instructions for the experiment (the "explicit instructional condition"). For the other half of the subjects, no advance warning was given, and the explicit face recognition test came as a surprise (the "on the fly instructional condition"). We assumed that if there were no differences across the two instructional conditions between the performance of subjects meeting some criterial level of accuracy on the explicit face recognition test, then it could safely be said that the subjects had followed the instructions as intended and had watched the visual display during study and test.

Method

Subjects

Subjects were 80 Indiana University undergraduates who participated in partial fulfillment of course requirements for Introductory Psychology. All subjects were native speakers of English, had normal hearing, and reported no history of speech or hearing disorders at the time of testing.

Stimulus Materials

STUDY AND TEST FOR WORD RECOGNITION: The stimuli consisted of 72 digitized movies of 8 talkers speaking isolated words. All of the items were taken from the Hoosier Multimodal Database (Sheffert et al., 1997). The average intelligibility of each word was 100% for all talkers used in the study, as indicated by the intelligibility data which accompanies the tokens in the HMD. All together, then, the total number of stimuli used in the experiment was 72 x 8 = 576. However, for each subject, only 72 tokens were ultimately viewed.

In order to provide the static picture of a talker's face for presentation, one movie from the set of 72 for each talker was selected as the "representative movie". The representative movie for each talker was a movie in which the second frame of the digitized sequence contained a picture of the talker with his/her mouth closed and which was not blurry. Thus, the second frame of each talker's representative movie was used as the static "face" stimulus for a particular talker for all subjects in Experiment 1.

Voice/word information was taken from the audio tracks of the 576 movies described above. For each subject, then, a semi-random pairing of faces and voices (preserving sex) was assigned prior to the commencement of the session. Then, for each subject, a randomly selected subset of half of the 72 words were taken to be "study" words. For each of these 36 words, one face-voice pairing was randomly assigned as the study presentation context.

For the test phase, the 36 "old" words were randomly assigned to one of four conditions. If a word was assigned to the (F+V+) condition, then it would be presented during test along with the same face-voice pairing with which it had been presented during study. If a word was assigned to the (F+v-) condition, then the face presented during study would be the same as the one with which the word was presented during study, but the voice would be a different, randomly assigned one. If a word was assigned to the (f-V+) condition, then the face would be a different randomly assigned one, but the voice would be the same as during study. Finally, if the word was assigned to the (f-v-) condition, then it would be presented with a totally new, randomly assigned face-voice pairing. However, this face-voice pairing *was* a pairing which had been seen consistently during study, just not with the specific word in question.

The 36 old words were then randomly mixed with the 36 new words to form the list of 72 test items. The 36 new words were randomly assigned to one of two possible conditions: one condition in which the randomly assigned face+voice pairing was one which had been a valid face+voice pairing during study, and one condition in which the face-voice pairing was not a valid one during study. Thus, the assignment of study words, face-voice pairings, the face-voice pairs which spoke each word and the conditions under which all words were tested were all randomly chosen for each subject.

TEST OF EXPLICIT FACE RECOGNITION: Stimuli for the explicit face recognition portion of this experiment were taken from another digitized set of video clips used as stimuli in a previous study (see Pisoni et al., 1995). Four of the talkers in the HMD were also talkers in this new database; "old" items in the explicit face recognition test were still frames taken from video clips of the four talkers who served as talkers in both databases. "New" items were still frames of four other video clips from the Pisoni et al. (1995) database. Thus, all the stimuli for the explicit face recognition test were drawn from the same set of digitized images. This set of eight explicit face stimuli were used for all subjects in all of the experiments reported here.

Apparatus

All stimuli were stored on an Pinnacle Micro 4.6 GB optical disk. A control program running on a Macintosh PowerPC 8100/100 assigned all random variables and presented stimuli according to this assignation. Visual stimuli were presented on a 17" Apple Multiple Scan 17 Display color monitor, controlled by a Radius video board, while auditory stimuli were presented over BeyerDynamic DT100 headphones calibrated to 74 dB SPL.

Procedure

Every other subject was assigned to one of the two instructional manipulations. For subjects in the "explicit" instructional condition, instructions were given which read as follows:

This experiment will consist of three parts.

- The first part is called the study phase. On each trial during this part, you will be shown a picture of a face on the computer monitor. At the same time, you will hear a word spoken over your earphones. Do your best to pay attention to both the word that is spoken and the face which is being shown, because you will be tested on them later.

- The second part is called the word recognition phase. On each trial, you will be shown a picture of a face and you will hear a word being spoken. After you hear the word, you should determine whether you have heard this word already in the study phase. You will then be required to indicate your response on the button box in front of you. Press the OLD button if you think you have already heard this word before. Press the NEW button if you think you have not. You should try to do this part as quickly as possible without sacrificing accuracy. Remember, you are judging whether the word you heard is "OLD" or "NEW."

- The last part will be a test on the faces you've seen. On each trial, you will see a picture of a face. Some of these will be faces you already saw in the previous phases, and some will be new. On each trial, your task is to determine whether you have already seen the face and indicate your response on the button box in front of you. You should press the OLD button if you think you have already seen this face. Similarly, you should press the

> NEW button if you think you have not seen the face before in the experiment. You should try to do this part as quickly as possible without sacrificing accuracy.

Subjects in the "on the fly" instructional condition, read the following instructions:
This experiment will consist of three parts. Instructions for each part will be posted on the computer screen before the beginning of each phase.

All subjects were given on-screen instructions before the commencement of each sub-section of the procedure.

During the test phase of the experiment, "old" or "new" responses for either words or faces, as the case may be, were collected by means of a button-box attached to a Strawberry Tree card. Responses, along with the parameters for a given trial, were recorded in data files for later analysis.

Results

In order for a subject to be included in the final data analysis, a score of 87.5% or better on the explicit face recognition test was necessary. The data from 36 subjects were eliminated in this way. In all, 23 subjects for each instructional condition were included in the final analysis.

Figure 1 shows the average d' scores for recognition of *words* in each testing condition for Experiment 1, as a function of Instructional Condition. This measure was chosen to assess the discriminability of old and new items in memory, consistent with previous approaches to the study of recognition memory (Banks, 1970; Egan, 1958; Lockhart & Murdock, 1970; Parks, 1966; Pollack, 1959; Shepard, 1967). The d' measure is advantageous with respect to measures of performance such as percent correct because it takes into account all the data collected for a given session. While accuracy measures only present data for responses to "old" items, d' is calculated using responses to both "old" and "new" items. Table 1 shows the corresponding Hits and False Alarm rates used in the calculation of d' scores for the various experimental conditions.

Insert Figure 1 about here.

It should be noted that responses to new word items cannot be split into the same number of conditions as responses to old word items. While old word items could occur in one of four contexts (F+V+, F+v-, f-V+, and f-v-), new words could only occur in one of two contexts (studied face+voice pairing, or non-studied face+voice pairing). The calculation of d' scores was carried out such that the False Alarm rate used for the F+V+ and f-v- conditions was the rate at which, in the context of a *previously studied* face+voice pairing, subjects responded "old" to new words. Similarly, the False Alarm rate used for the calculation of d' scores in the F+v- and f-V+ conditions was the rate at which subjects responded "old" to new words presented in the context of face+voice pairings which had *not* been previously studied.

In other words, two different False Alarm rates were used in the calculation of d' scores. One False Alarm rate was used for the calculation of d' scores for stimuli which were presented with *experimentally valid* face+voice pairings. The other False Alarm rate was used for the calculation of d' scores for stimuli which were presented in face+voice pairings which *had not been studied*.

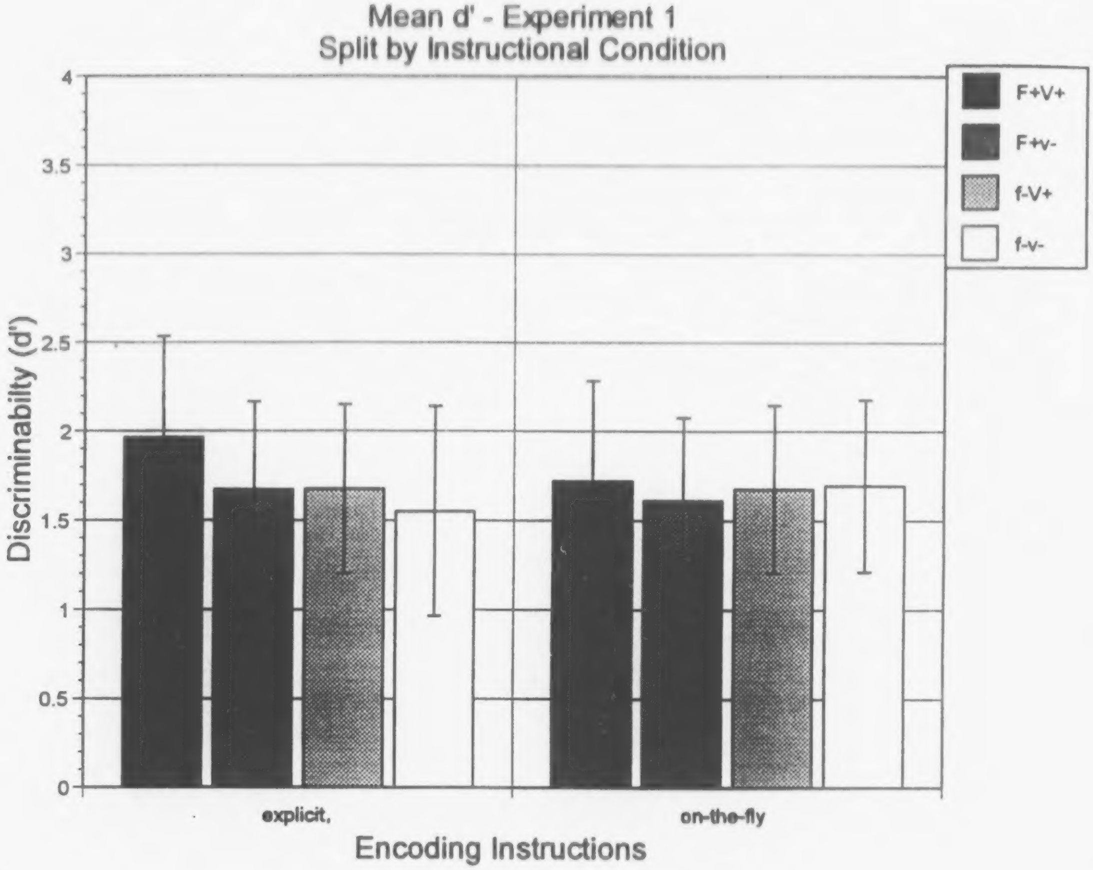


Figure 1: Average d' scores for all four conditions in Experiment 1 as a function of instructional condition.

178

Table 1

Average Hits and False Alarms for all Conditions in Experiment 1, Split by Instructional Condition.

			Encoding	Instructions			
explicit face+voice	average FAs	condition	average Hits	on the fly face+voice	average FAs	condition	average Hits
non-studied	3.30	F+V+	8.00	non-studied	4.30	F+V+	6.87
		f-v-	6.35			f-v-	7.35
studied	3.70	F+v-	7.22	studied	3.61	F+v-	6.57
		f-V+	6.35			f-V+	6.61

The left panel of Figure 1 shows the average d' scores from subjects in the "explicit" instructional condition for the four experimental conditions, (F+V+), (F+v-), (f-V+) and (f-v-). The right panel of Figure 1 shows the analogous scores taken from subjects in the "on the fly" instructional condition. As can be seen from the graph of "on the fly" scores, there were virtually no differences between discriminability scores in each of the experimental conditions. However, the left panel reveals a notable difference between discriminability in the (F+V+) condition and the rest of the conditions. Similarly, there seems to be at least a numerical advantage to discriminability in the conditions (F+v-) and (f-V+) relative to the (f-v-) condition.

A repeated measures ANOVA was conducted on the d' scores for Experiment 1 with Face Context (new or old) and Voice Context (new or old) as the repeated measures and Instructional Condition ("explicit" or "on the fly") as a between subjects variable. This analysis revealed a significant effect of Voice Context, F(1,44) = 3.686, p = 0.061. Across instructional conditions, subjects were better able to distinguish old from new words when the old words were presented in the context of the voice with which that word was originally studied.

In addition, the interaction between Face Context and Instructional Condition approached significance, F(1,44) = 2.849, p = 0.099. Figure 2 displays the Face Context x Instructional Condition interaction for d' scores in Experiment 1. The left panel of Figure 2 shows the discriminability scores for subjects in the "explicit" instructional condition; the left side shows scores for subjects in the "on the fly" instructional condition. Each bar represents the average discriminability of items, collapsed across voice condition. Thus, the "FaceOLD" bar on both sides of the figure represents the average discriminability for items in the (F+V+) and (F+v-) conditions. Similarly, the "FaceNEW" bar for each panel of the figure represents the average discriminability for items in the (f-V+) and (f-v-) conditions. A probe of this interaction through a 2 (Face Context) x 2 (Voice Context) repeated measures ANOVA split by Instructional Condition revealed that this effect was due to a significant main effect of Face Context in the "explicit" instructional condition F(1,22) = 4.831, p = 0.039, but not in the "on the fly" instructional condition, F(1,22) = 0.048, n.s. Thus, for subjects in the "explicit" instructional condition only, the distinction between old and new items was significantly greater when those items were presented in the context of the face with which they were originally presented.

Insert Figure 2 about here.

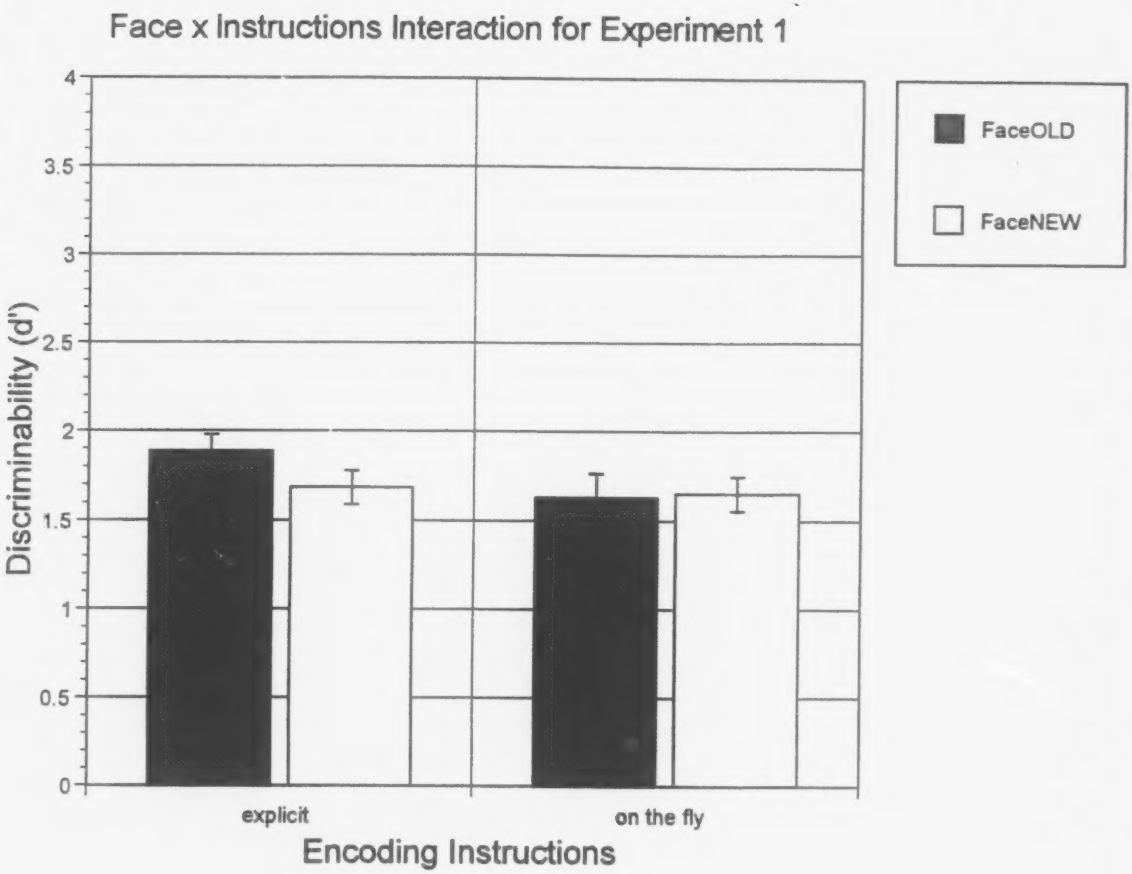


Figure 2: Average d' scores for test conditions in which the face was old or new as a function of instructional condition. The FaceOLD bar represents the average d' scores for test items in the (F+V+) and (F+v-) conditions. The FaceNEW bar represents the average d' scores for test items in the (f-V+) and (f-v-) conditions.

Discussion

The results from Experiment 1 replicated and extended the major findings of Kato et al. (1995) in accordance with our original predictions. We observed an effect of modality context on the recognition of words. Old words presented in the context of an old voice were more discriminable from new words than old words presented in a different voice context. This effect replicates the voice repetition effect found in previous studies (Goldinger, 1995; Palmeri, Goldinger, & Pisoni, 1991) and establishes the validity of the current experimental procedures.

Furthermore, a significant effect of Face Context on the recognition of words was found in the "explicit" instructional condition. In retrospect, it seems likely that the "explicit" instructional condition may represent a task which is more closely tied to the procedures used by Kato et al. (1995) than originally expected, because it may result in an encoding strategy for face information which allows for the "explicit" recall of face information at some later point. There is evidence to show that the encoding of information for "explicit" and implicit retrieval may be dissociated from one another (Schacter & Church, 1992) In light of these findings, it may be that the instructional manipulation was more fruitful than originally anticipated. To the extent that the instructions given in the "explicit" instructional condition result in a strategy of "explicit" encoding of face information *and* to the extent that the instructions given in the "on the fly" condition do not, it can be said that the instructional manipulation represents a direct test of one of our hypotheses. Namely, it tests the assertion that an "explicit" task will not foster the need for multimodal encoding since the demands of such a task do not require it. In other words, asking a subject to explicitly recall information from one modality may only invoke the information in another modality presented at study due to simple co-occurrence. By contrast, an implicit task requires encoding of an underlying event, and as such, may make use of all information which was relevant to the instantiation of the event being recognized.

In light of this interpretation of the instructional manipulation, an explanation for the Face Context x Instructional Condition interaction may be seen. We claim that conclusions about the integration of audiovisual information in memory may not be drawn when the visual information used to test this integration is static, because such displays eliminate any natural relationships which may exist between the visual and auditory information and thus obviate any perceptual processes which may serve to integrate multimodal inputs. However, we do not deny that such displays may serve as effective retrieval cues during recognition (as in Kato et al., 1995; Legge et al., 1984). Indeed, our hypothesis implies that it is the *relationship* between information in the various input modalities which effects their link in memory. Thus, if some experimentally induced manipulation lends credence to the relationship between information perceived through disparate sensory modalities, then that relationship should be encoded, and thereafter used during recognition processes. In other words, explicitly alerting subjects to the fact that a static face is related to an associated speech event - either through instructions for an explicit task, as in Kato et al. (1995) and Legge et al. (1984), or through instructions similar to our own - will result in a bond in memory between the information coming from either modality. The fact that static faces were used as an effective retrieval cue for recognition of words in the "explicit" instructional condition, but *not* in the "on the fly" condition, indicates that, under normal circumstances, there is no integral encoding of static visual information with simultaneously presented speech. With manipulation of specific task demands, though, the memory system may forge an arbitrary link where no natural one occurs.

In summary, Experiment 1 can still be considered a extension of Kato et al. (1995) because it shows that using single words spoken in isolation as the carrier of voice information (as opposed to sentences) and using an implicit memory paradigm (as opposed to an explicit one) does not change the basic finding that audiovisual relationships do exist in memory and can be constructed as a result of arbitrary pairings of spoken words with static faces (Kato et al., 1995; Legge, Grosmann, & Pieper, 1984); however, some sort of experimental manipulation must be present to force the bond if the visual information is static.

Experiment 1 therefore confirms that, through experimental manipulation, it is possible to induce in subjects a strategy for encoding events in recognition memory which in some way links co-occurring information from disparate modalities. The results, however, seem to imply that this cross-modal link in memory is due to purely arbitrary reasons: faces can only be used as effective retrieval cues for the recognition of speech events when subjects are explicitly instructed as to their relevance to the task at hand. It may be, then, that *any* arbitrary co-occurring visual information could be encoded in the same way, given the right conditions. Thus, these results do not necessarily add to our understanding of speech perception and spoken word recognition in audiovisual environments. Experiment 2 was designed to test whether the findings from Experiment 1 with static faces generalize to results which might be found with *any* visual stimulus. If so, then conclusions may not be drawn from static stimuli concerning the integration of face and voice information in representations of speech in memory.

Experiment 2

The task for subjects in Experiment 2 was the same as in Experiment 1; once again, an explicit face recognition test was used as a criterion for entry into subsequent data analyses, and the between-subjects instructional condition was again used. However, in this experiment, the faces were rotated 180° and presented upside down. We reasoned that, because the pairing of voices and faces was arbitrary in the Kato et al. (1995) experiment, then *any* arbitrary pairing of any visual stimulus with voice information should produce essentially the same recognition memory effects.

Several earlier studies have shown that recognition of faces is impaired by upside down presentation (Valentine, 1988; Yin, 1969). By presenting the faces upside down, we controlled for visual complexity of the visual stimulus, while simultaneously reducing the cue value of the facial information. One caveat to this approach, however, is that upside down faces are unfamiliar to subjects.

The aim of this experiment was to show that, while experimentally derived conditions may well explore a link between unimodal representations, they may not be able to shed light on the matter of natural multimodal integration. Showing that static faces can be used as a retrieval cue for recognition of voices (Kato et al., 1995; Legge et al., 1984) may be no better than showing that *any* arbitrary concurrent other mode stimulus can be used as a retrieval cue, and, therefore, cannot provide new insights into the matter of audiovisual speech representations in memory.

Method

Subjects

Subjects were 40 Indiana University undergraduates who participated in this experiment as partial fulfillment of course requirements for Introductory Psychology. All subjects were native speakers of English, had normal hearing, and reported no history of speech or hearing disorders at the time of testing.

Stimulus Materials

STUDY AND TEST FOR WORD RECOGNITION: The stimuli in Experiment 2 were identical to those used in Experiment 1 in all but one respect: in Experiment 2, the static face part of each face+voice pairing was rotated and presented upside down. This was accomplished by taking the "representative movie" (as described above) and passing it through a 180 degree rotation matrix in Adobe Premier v.4.0. The output rotated movie was then used as the "representative movie" for a particular talker across subjects.

TEST OF EXPLICIT FACE RECOGNITION: Stimuli for the explicit face recognition portion of this experiment were the same as in Experiment 1. This includes their orientation.

Apparatus

Experiment 2 was carried out using the exact same apparatus as Experiment 1.

Procedure

Procedures for Experiment 2 were analogous to those in Experiment 1, except that whenever it was necessary to mention the nature of the face information, subjects were told that the faces would be presented upside down on their video screens.

Once again, the assignment of study words, face-voice pairings, the face-voice pairs which spoke each word and the conditions under which all words were tested were all randomly assigned for each subject.

During the test phase of the experiment, "old" or "new" responses for either words or faces, as the case may be, were collected by a button-box interfaced with a Strawberry Tree card. Responses, along with the parameters for a given trial, were recorded in data files for later analysis.

Results

Due to extremely poor performance on the explicit face recognition task, very few subjects were able to meet the criterion for entry into the analysis of 87.5% accuracy. The criterion was therefore relaxed to 75% accuracy so that statistical analyses could be carried out. Even with the weaker criterion, only 9 subjects from the "on the fly" instructional condition and 10 subjects from the "explicit" instructional condition could be included in the final analysis.

Insert Figure 3 about here.

Figure 3 shows the average d' scores in each testing condition for Experiment 2, separated as a function of Instructional Condition. Figure 3 follows the same format as Figure 1, with the left panel representing average discriminability scores in each of the experimental conditions (F+V+), (F+v-), (f-V+) and (f-v-) from subjects in the "explicit" instructional condition. The right panel of Figure 3 shows the scores from subjects in the "on the fly" instructional condition. Table 2 shows the corresponding Hits and False Alarm rates used in the calculation of d' scores for the various experimental conditions. As before, the calculation of d' scores was carried out such that the False Alarm rate used for the (F+V+) and (f-v-) conditions was the rate at which subjects responded "old" to new words presented in the context of a previously studied face+voice pairing. Similarly, the False Alarm rate used for the calculation of d' scores in the (F+v-) and (f-V+) conditions was the rate at which subjects responded "old" to new words presented in the context of face+voice pairings which had not previously been studied together.

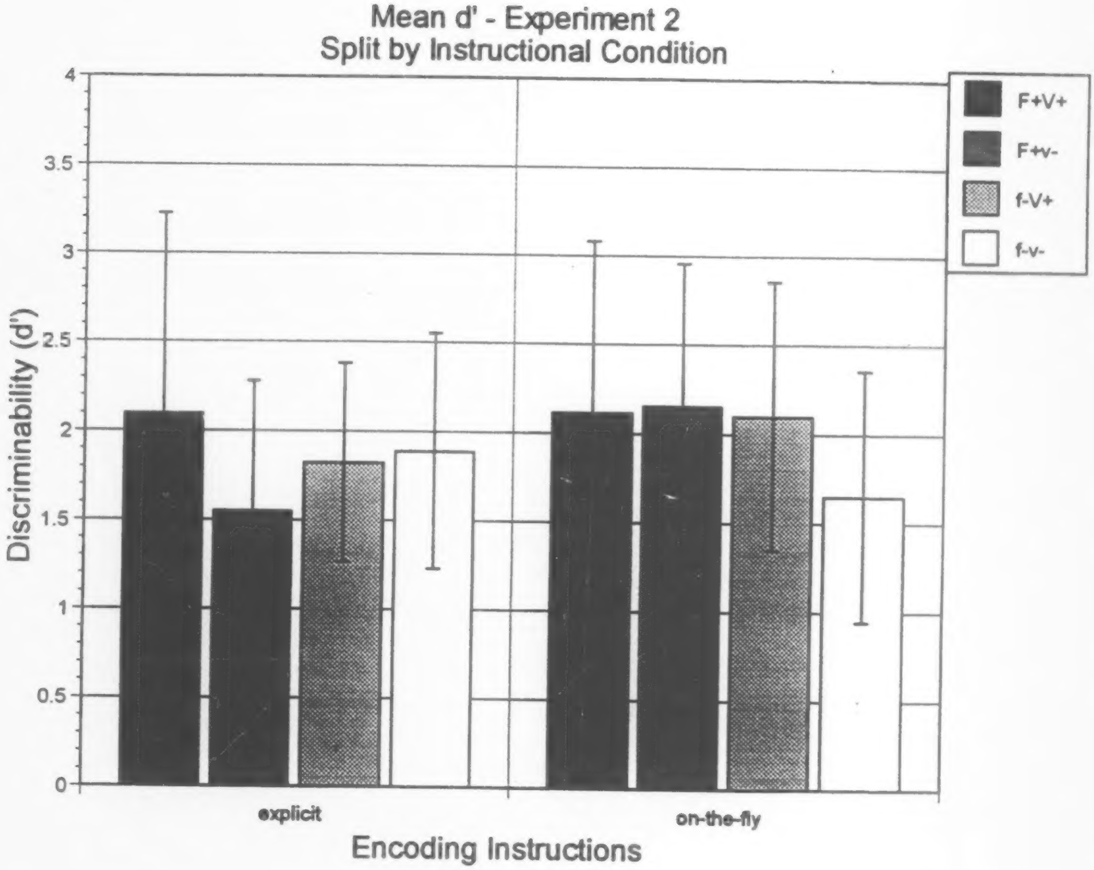


Figure 3: Average d' scores for all four conditions in Experiment 2, as a function of instructional condition.

Table 2

Average Hits and False Alarms for all Conditions in Experiment 2, Split by Instructional Condition.

			Encoding	Instructions			
explicit				on the fly			
face+voice	average FAs	condition	average Hits	face+voice	average FAs	condition	average Hits
non-studied	3.70	F+V+	7.60	non-studied	4.78	F+V+	5.89
		f-v-	7.60			f-v-	5.67
studied	4.20	F+v-	6.50	studied	5.22	F+v-	7.00
		f-V+	7.10			f-V+	6.67

A repeated measures ANOVA was conducted on the d' scores for Experiment 2 with Face Context (new or old) and Voice Context (new or old) as the repeated measures and Instructional Condition ("explicit" or "on the fly") as a between subjects variable. The results of this analysis revealed no significant main effects or interactions.

Discussion

It seems at first glance as though Experiment 2 did not confirm the hypothesis it was designed to test. This experiment was designed to show that *any* visual stimulus which was arbitrarily paired with voice information could act as an effective cue for retrieval during recognition of a speech event, and therefore could not speak to the issue of audiovisually integrated representations in memory. In order to test this hypothesis, an upside down face was chosen as a suitable visual stimulus, since such a stimulus would, when compared to performance with static faces, control for the complexity of the visual image while simultaneously eliminating any cues to the "face-ness" of the stimulus (Valentine, 1988; Yin, 1969).

However, in retrospect, given the outcome of this study, it seems as though this particular visual stimulus may not have been an appropriate one; given the previous findings showing the difficulty subjects have with explicitly recognizing upside down faces, it is not surprising that performance on the explicit face recognition task (with upright faces) was so poor. This poor performance is potentially confounding in several respects.

First, despite the relaxation of the entry criterion, the data from very few subjects were actually included in the final analysis. As such, the low power associated with each statistical test may have acted to obscure any underlying trends.

Second, and perhaps more importantly, it may be that the explicit face recognition task did not test what it was designed to test in the context of Experiment 2. Originally, the explicit face recognition test was added to the design of the experiment so that we could have an objective measure of how well subjects followed the instructions to pay attention to both visual and auditory information during study. As such, we assumed that the measure should remain constant across experiments, so that parallels could be drawn between the data collected in each. However, it now can be seen that, in the context of Experiment 2, the explicit face recognition test also measured another variable - that is, it assessed how well subjects could transfer their knowledge of the upside down faces in the study and word test phases to rightside up faces.

As stated above, such a transfer of knowledge has already been shown to be poor and was the very reason we selected upside down faces as visual stimuli in the first place.

There are two ways in which the design of Experiment 2 could be altered to eliminate this confound and simultaneously test the original hypothesis. First, the explicit face recognition phase could be altered such that the stimuli included in it are also upside down. This would eliminate the need for a transfer of knowledge about upside down faces to recognition of upright faces. In this way, the explicit recognition task would be more clearly comparable to that in Experiment 1.

However, it may be that the inappropriateness of the stimuli in Experiment 2 goes further than that. Subjects are, in general, extremely practiced at recognizing upright faces and have very little experience recognizing upside down ones. It therefore may be necessary to use a different type of visual stimulus for Experiment 2, such as animal faces, geometric shapes, houses, cars, or patches of color. Because these visual stimuli are usually well-learned in the experience of most subjects, their recognition may be more directly related to the recognition of static faces.

Still, the results from Experiments 1 and 2 make one thing perfectly clear: under normal circumstances, static visual displays, regardless of their content, are *not* integrated in memory with information about a simultaneously presented speech event. For the subjects in the "on the fly" instructional condition of both experiments, visual information presented concurrently was not used as an effective retrieval cue for recognition of a speech event. However, in Experiment 1, explicit instructions as to the importance of the visual stimulus did forge a connection in memory between visual and auditory information. These explicit instructions may well have forced a connection between visual and auditory information in Experiment 2 also, but, given subjects' difficulty in encoding upside down faces, the encoding of upside down faces may have been insufficient to foster the use of such information as a retrieval cue. In any event, Experiments 1 and 2 show that arbitrary visual information may not serve as an effective retrieval cue for speech events in the absence of sufficient experimentally induced bias.

Can visual information act as an effective retrieval cue when it is *not* arbitrarily paired with the speech event? A dynamic visual display of a talker speaking contains information which is not arbitrarily tied to the underlying speech event. Indeed, the auditory specification of a speech event is *lawfully tied* to its articulation. Our hypothesis, therefore, was that subjects would be able to use a dynamic optic display as an effective retrieval cue during recognition of a speech event, since this type of display provides information concerning the speech event during encoding. In other words, we postulated that any transfer to memory that takes place during speech perception will *automatically* encode any information which is relevant to the event being encoded; since dynamic optical displays of articulation are informative about a speech event, then it follows that visual information will also be encoded and later used as an effective retrieval cue during recognition. As such, Experiment 3 was designed to test the integration of multimodal information in memory for speech using dynamic visual displays of talkers uttering the stimuli.

Experiment 3

The task for subjects in Experiment 3 was the same as in the previous two experiments. However, this time, the visual stimuli were dynamic video clips of talkers articulating the words. As in the previous experiments, the explicit face recognition test and the instructional condition were used. We reasoned that because intermodal relationships in the sensory information are preserved in dynamic visual displays, the results from this experiment would provide evidence for the integration of audiovisual speech representations in memory. We expected that we would once again find evidence for the use of intermodal relationships in memory in the form of main effects of both face and voice. In addition, we hypothesized that overall levels of performance in Experiment 3 would be greater than in Experiment 1, due to increased ability of subjects to exploit the naturally occurring, lawful, intermodal relationships between faces and voices specifying speech events.

It is worth noting here that our post hoc interpretation of the instructional manipulation in Experiment 1 has ramifications for the outcome of Experiment 3, especially when considered in the light of our experimental hypothesis. Our post-hoc interpretation of the instructional manipulation assumes that the "explicit" instructional manipulation causes a change in the strategies utilized by subjects in their encoding of face information, resulting in an experimentally induced arbitrary association between face and voice information in memory. Thus, in Experiment 1, static faces could only be used as effective retrieval cues during recognition of speech events when subjects were *explicitly* told that faces were important in the experiment and that they would be tested at the end of the experiment. Under normal circumstances, a static face would *not* be linked in the encoding of a speech event because its non-dynamic nature serves to eliminate any information which could specify the phonetic context. These hypotheses were supported by the findings of Experiment 1.

The prediction made by these hypotheses, however, is that if the visual display contained information specifying its relation to the acoustic signal (i.e., if the visual display provides dynamic information about the articulation of the speech being heard) then a link between the auditory and visual information should be forged *in the absence of explicit instructions to do so.* In other words, the integrated encoding of dynamic visual information in memory for speech events should be mandatory, since a dynamic optic display of a talker's articulation will specify the nature of the linguistic message. As such, both face and voice context should show an effect on the recognition of words in the "on the fly" instructional condition, when only the information that is naturally encoded in a mandatory fashion will implicitly affect recognition performance.

The effects of dynamic visual information on recognition for subjects in the "explicit" instructional condition are not directly predictable from the hypotheses, but two possibilities exist. The first alternative is that explicit instructions to associate face and voice information will only serve to strengthen or reinforce the bond between multimodal information in memory, and as such, subjects in the "explicit" instructional condition should show the same pattern of results as in Experiment 1, albeit with higher recognition scores. The second alternative is that the *experimentally induced* bond between cross-modal information in memory may serve to counteract the effects of the *naturally occurring and lawful* bond which relates the information in both modalities to a common underlying perceptual event. Thus, the second alternative predicts that the effects of face context in the recognition of words will be eliminated by explicit instructions during the dynamic experiment.

Method

Subjects

Subjects were 40 Indiana University undergraduates who participated in partial fulfillment of course requirements for Introductory Psychology. All subjects were native speakers of English, had normal hearing, and reported no history of speech or hearing disorders at the time of testing.

Stimulus Materials

STUDY AND TEST FOR WORD RECOGNITION: The stimuli used in Experiment 3 were taken from the same digital database as those for Experiments 1 and 2. However, face context information was now provided by the entire video track that corresponded to the word being presented. In addition, the digitized movie clips were truncated such that the number of frames to the onset of speech was equal for all movies showing the different talkers speaking a given word. Thus, face+voice pairings were naturally occurring ones and were not randomly assigned.

Another computer program was written to trace out the course of each subject's experimental session. This "plotting" program randomly assigned test items to serve as old or new words, randomly determined the order in which old and new words were presented during the test phase, and randomly determined the audiovisual context in which a word could occur during test. The program then determined which stimuli were composed of face+voice pairings where the face and voice came from different people. For these stimuli, the appropriate video track was dubbed on to the corresponding audio track by four steps. First, the header file for the movie which contained the necessary audio track was obtained. Second, a new digital movie structure was created using the audio information from the appropriate audio track. Third, the video track pointer for the new movie was changed so that it pointed to the appropriate video track. Fourth, the data were flattened and used to write a new "dubbed" movie, which was stored in a location which was accessible later. Because the lead-in time for all movies which could possibly be dubbed together was equal, synchronization of the video and audio tracks was accomplished by default, with a maximum possible error of 16.5 ms, half the duration of one movie frame. It should be noted that running *correct* face+voice pairing movies through this same process would have resulted in output movies identical to those input, and as such, only movies that needed to be dubbed were actually processed in this manner, in order to save time and space.

The output of the "plotting" program was used as input to a modified version of the control program used in Experiments 1 and 2. The modified version simply played out the movies (study, dubbed and non-dubbed) in the order specified by the plotting program. The control program also displayed on-line instructions and collected responses, as before.

TEST OF EXPLICIT FACE RECOGNITION: Stimuli for the explicit face recognition portion of this experiment were identical to those used in Experiments 1 and 2 (i.e., they were static faces in their upright orientation).

Apparatus

A Macintosh PowerPC 8100/100 was used to run both the plotting program and the control program. Due to the increased demand for fast data transfer between the drive where the movies were stored and the video processing board, a ProDirect 9.6 GB hard drive with an Ultra-SCSI connection was used to store the stimuli. As in Experiments 1 and 2, the video track of each stimulus movie was presented on a 17" Apple Multiple Scan 17 Display color monitor, controlled by a Radius video board, while the audio track of each stimulus movie was presented over Beyer dynamic DT100 headphones calibrated to 74 dB SPL.

Procedure

Procedures for Experiment 3 were identical to those in Experiment 1, except that subjects were always told that the stimuli they would be viewing were movies.

As in the previous experiments, "old" or "new" responses for either words or faces, as the case may be, were collected by a button-box interfaced with a Strawberry Tree card. Responses, along with the parameters for a given trial, were recorded in text files for later analysis.

Results

In order for a subject to be included in the final data analysis, a score of 87.5% or better on the explicit face recognition memory test was necessary; the final analysis included 15 subjects from the "on the fly" instructional condition, while 17 subjects were included from the "explicit" instructional condition.

Insert Figure 4 about here.

Figure 4 shows the average d' scores in each testing condition for Experiment 3, separated as a function of Instructional Condition. The left side of the figure shows scores obtained in each of the four within-subjects conditions from subjects participating in the "explicit" instructional condition. The right side shows the same scores obtained from subjects in the "on the fly" instructional condition. Table 3 shows the corresponding Hits and False Alarm rates used in the calculation of d' scores for the various experimental conditions. As before, the calculation of d' scores was carried out such that the False Alarm rate used for the (F+V+) and (f-v-) conditions was the rate at which subjects responded "old" to new words presented in the context of a previously studied face+voice pairing (which, in the case of Experiment 3, was a naturally occurring face+voice pair). Similarly, the False Alarm rate used for the calculation of d' scores in the (F+v-) and (f-V+) conditions was the rate at which subjects responded "old" to new words presented in the context of face+voice pairings which had not been previously studied, and which do not occur naturally.

Table 3

Average Hits and False Alarms for all Conditions in Experiment 3, Split by Instructional Condition.

			Encoding	Instructions			
explicit				on the fly			
face+voice	average FAs	condition	average Hits	face+voice	average FAs	condition	average Hits
non-studied	3.06	F+V+	7.35	non-studied	2.47	F+V+	7.47
		f-v-	6.47			f-v-	6.40
studied	3.53	F+v-	6.94	studied	2.53	F+v-	7.47
		f-V+	7.88			f-V+	7.47

A repeated measures ANOVA was conducted on the d' scores for Experiment 3 with Face Context (new or old) and Voice Context (new or old) as the repeated measures and Instructional Condition ("explicit" or "on the fly") as a between subjects variable. This analysis revealed a significant effect of Voice Context $F(1,30) = 11.125$, $p = 0.002$. Across both instructional conditions, subjects were better able to discriminate old from new words when the old words were presented in the context of the voice with which that word was studied. In addition, the interaction between Face Context and Instructional Condition was significant, $F(1, 30) = 4.977$, $p = 0.033$.

Figure 5 illustrates the Face Context x Instructional Condition interaction for d' scores in Experiment 3. The left panel of Figure 5 shows the discriminability scores for subjects in the "explicit"

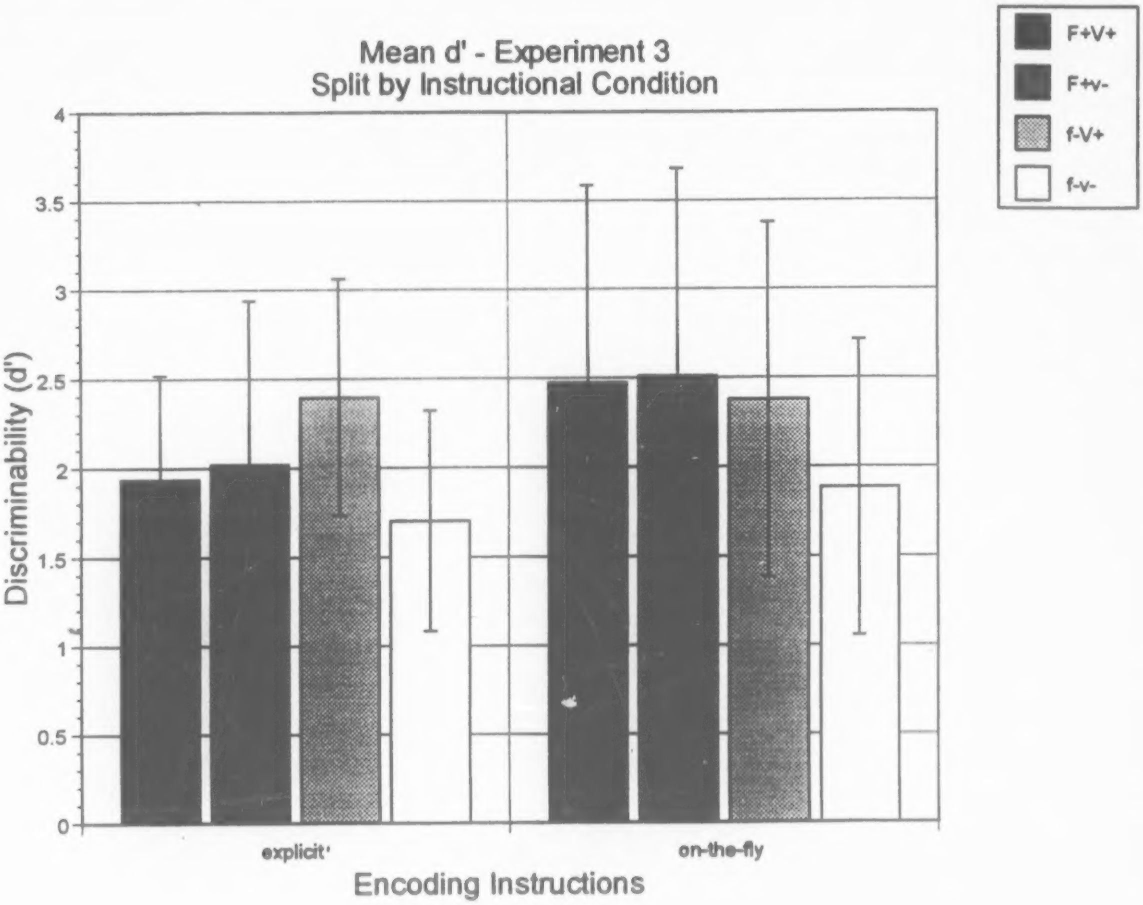


Figure 4

Figure 4: Average d' scores for all four conditions in Experiment 3, as a function of instructional condition.

instructional condition; the left side shows scores for subjects in the "on the fly" instructional condition. As in Figure 2, each bar represents the average discriminability of items, collapsed across voice condition. A probe of this interaction using a 2 (Face Context) x 2 (Voice Context) repeated measures ANOVA split by Instructional Condition revealed that the interaction was due to a highly significant main effect of Face Context in the "on the fly" instructional condition only, F(1,14) = 6.827, p = 0.020, but not in the "explicit" instructional condition, F(1,16) = 0.280, N.S. Thus, for subjects in the "on the fly" instructional condition, the distinction between old and new word items was significantly greater when those items were presented in the context of the face with which they were originally studied.

Insert Figure 5 about here.

Across Experiment Analysis

We also predicted that overall levels of performance would be higher in Experiment 3 when compared to Experiment 1 because subjects are able to make use of the naturally occurring and lawful bonds between information presented in the two separate modalities. In order to test this hypothesis, the data from both Experiments 1 and 3 were submitted to a repeated measures ANOVA with Face Context (new or old) and Voice Context (new or old) as the repeated measures and Instructional Condition ("explicit" or "on the fly") and Experiment (1 or 3) as between subjects factors. Figure 6 illustrates all the data which was analyzed in this way. This figure is a restructuring of the data presented in Figures 1 and 4, so that the scores for each condition may better be compared across experiments. Figure 6a shows the scores from all subjects in the "explicit" instructional condition; Figure 6b shows the scores from all subjects in the "on the fly" instructional condition. The light bars represent scores obtained from subjects in Experiment 1 with static faces, while the dark bars represent scores obtained from subjects in Experiment 3 with dynamic faces.

Insert Figure 6 about here.

As can be seen from Figures 6a and 6b, scores were generally higher in Experiment 3 than they were in Experiment 1. Indeed, a significant main effect of Experiment was found, F(1,74) = 11.987, p = .001, indicating that subjects in Experiment 3 with dynamic visual stimuli had better recognition scores than those in Experiment 1 with static visual stimuli. In addition, a highly significant main effect of Voice Context was found, F(1,74) = 14.442, p < 0.0009, whereas there was no interaction between Voice Context and Experiment, F(1,74) = 1.827, n.s. This demonstrates that the voice effect was consistent across experiments. In other words, both Experiments 1 and 3 replicated previous findings showing that word recognition is affected by voice context (Goldinger, 1995; Palmeri et al., 1991); old words were more easily discriminated from new words when they were presented during test with the same voice as in test.

Finally, we observed a significant interaction between Face Context, Instructional Condition, and Experiment, F(1, 74) = 8.351, p = .005. Figure 7 shows the interaction of Face Context and Experiment as a function of Instructional Condition. The graph in Figure 7 is a recombination of the data in Figures 2 and 5, reorganized to facilitate comparison across experiments. Thus, the left panel shows scores, collapsed across voice condition, for subjects who participated in the "on the fly" instructional condition for Experiments 1 and 3, while the right panel shows scores obtained from the "explicit" condition. Since this interaction is not easily interpretable, the data were split into two groups based on Instructional Condition,

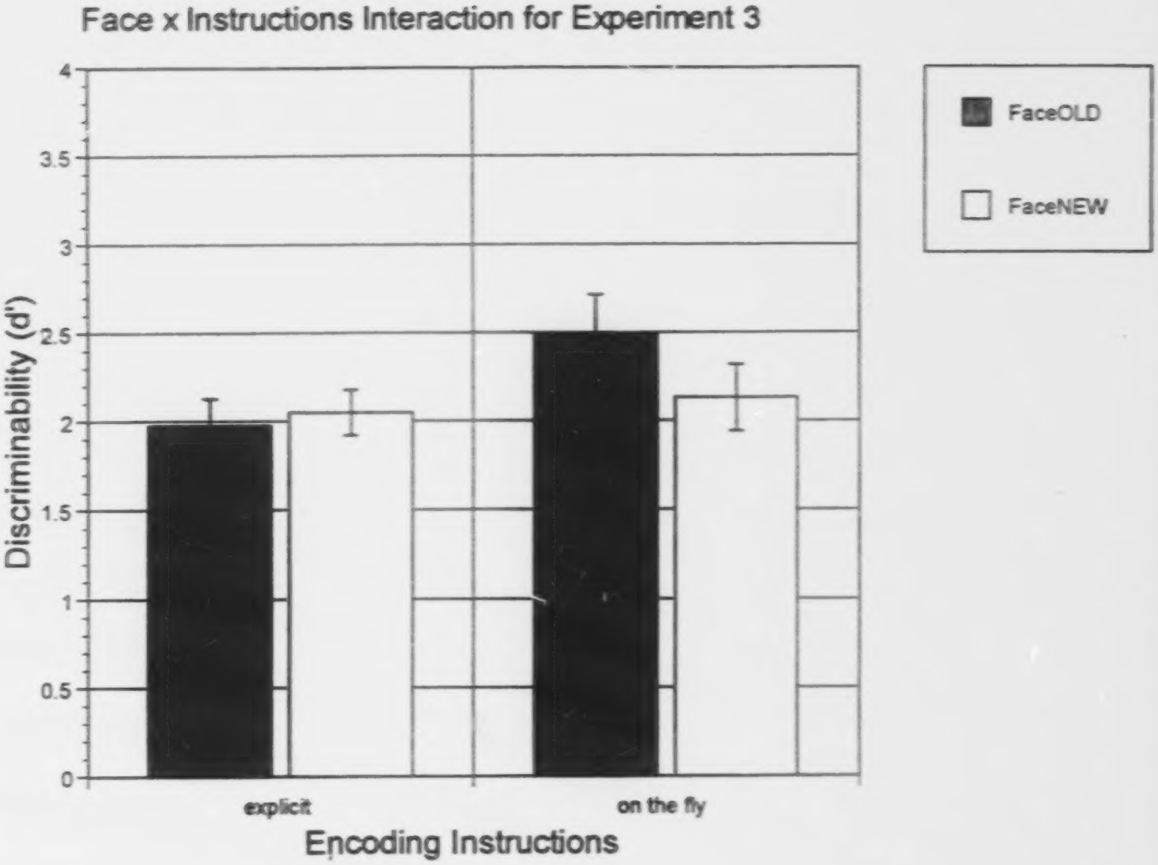
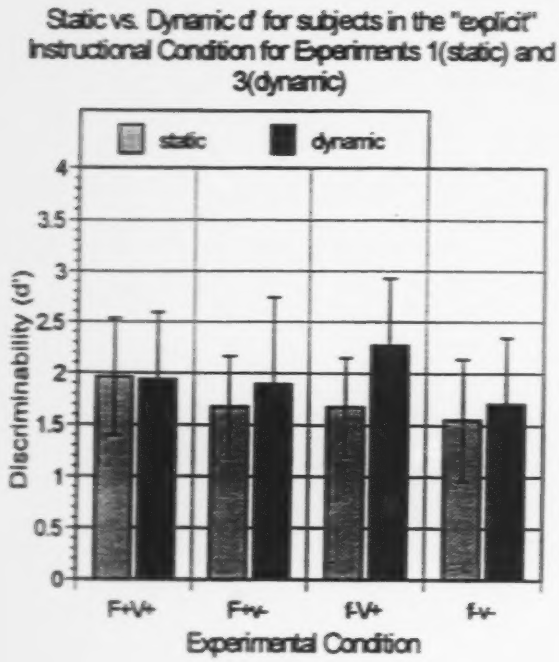


Figure 5: Average d' scores for test conditions in which the face was old or new as a function of instructional condition. The FaceOLD bar represents the average d' scores for test items in the (F+V+) and (F+v-) conditions. The FaceNEW bar represents the average d' scores for test items in the (f-V+) and (f-v-) conditions.


Static vs. Dynamic d' for subjects in the "explicit" Instructional Condition for Experiments 1(static) and 3(dynamic)

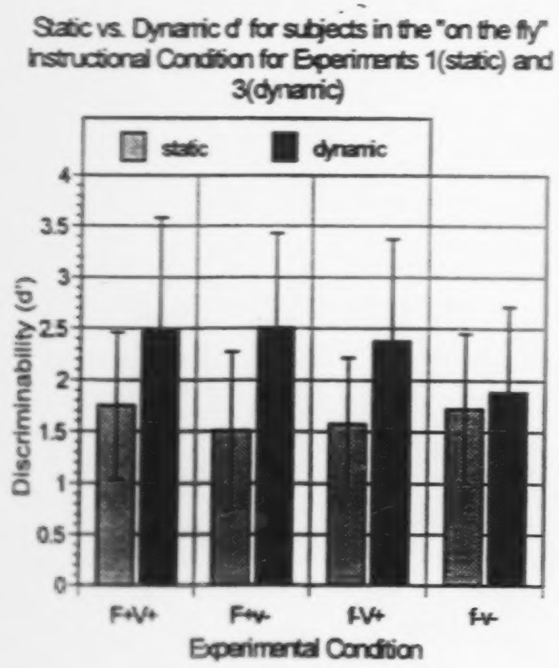

Static vs. Dynamic d' for subjects in the "on the fly" Instructional Condition for Experiments 1(static) and 3(dynamic)

Figure 6a: comparison of average d' scores for subjects in the "explicit" instructional condition of Experiments 1 and 3.

Figure 6b: comparison of average d' scores for subjects in the "on the fly" instructional condition of Experiments 1 and 3.

and each group's data were submitted to a repeated measures ANOVA, using Face Context and Voice Context as repeated measures and Experiment as a between subjects factor.

Insert Figure 7 about here.

The 2 (Face Context) x 2 (Voice Context) x 2 (Experiment) ANOVA for subjects in the "on the fly" instructional conditions of both experiments revealed a significant main effect of Experiment, $F(1,36) = 10.575$, $p = 0.002$: subjects' performance with "on the fly" instructions was always better in Experiment 3 (dynamic visual displays) than in Experiment 1. The ANOVA also revealed a significant Face Context by Experiment Interaction, $F(1,36) = 5.469$, $p = 0.025$. A post-hoc t-test confirmed that the source of this interaction was a difference in the effects of Face Context on the recognition scores for items presented in Experiment 3, $t(14) = 2.613$, $p = 0.02$. For subjects who were not explicitly instructed to attend to the faces in the experiment, Face Context only affected recognition scores when visual displays were dynamic. This finding can be interpreted as a form of implicit encoding of face information.

Analysis of the data from subjects in the "explicit" instructional condition also revealed a marginally significant interaction between Face Context and Experiment, $F(1,38) = 3.036$, $p = 0.09$. A post-hoc t-test confirmed that the source of this interaction was a difference in the effects of Face Context on the recognition scores for items presented in Experiment 1, $t(22) = 2.198$, $p = 0.039$: For subjects who were explicitly instructed that faces were important in the experiment, Face Context could only serve as an effective cue to recognition of words when visual displays were static.

In order to understand more fully the cross-experiment results, η^2 values were computed for those main effects of Face Context that were found to be significant. Since η^2 is taken to be the proportion of variance accounted for by a particular effect, it is useful as a measure of the relative magnitudes of different effects. For subjects in Experiment 1 (static visual displays) who were given the "explicit" instructions, the main effect of Face Context accounted for 18% of the total variance. For subjects in Experiment 3 (dynamic visual displays) who were given "on the fly" instructions, the main effect of Face Context accounted for 32.8% of the total variance, almost double that found in Experiment 1. Thus, for those subjects who were able to use face cues (given the experiment and instructional condition in which they participated), dynamic information was more useful than static information.

Discussion

Experiment 3 replicated the finding that same-voice repetition of test items can facilitate later recall and recognition (Goldinger, 1995; Palmeri et al., 1995); across instructional groups and face contexts, old items were recognized better when the word was presented in the context of the voice with which it was originally presented than when it was presented in the context of a new voice. It is safe to assume then, that our experimental procedure was a valid one, because we were able to consistently find a repetition effect.

As expected, the findings from Experiment 3 are consistent with the hypothesis that dynamic visual information is implicitly encoded in memory and can be used as an effective retrieval cue in the recognition of words. For subjects in the "on the fly" instructional condition, recognition memory for words was improved when those words were presented in the context of the dynamic, articulating face with which it was originally presented. This finding is consistent with the hypothesis that the integrative encoding of visual information in memory for speech is automatic and mandatory when the face information is dynamic

BEST COPY AVAILABLE

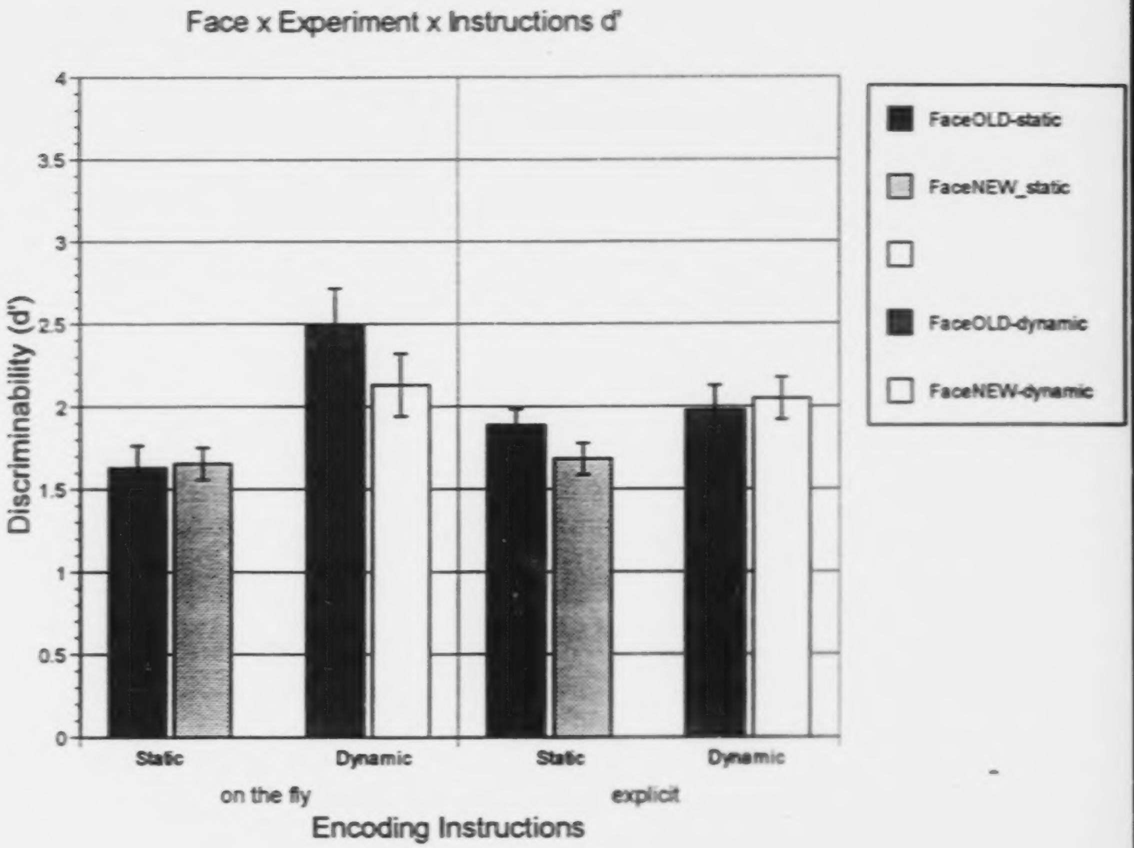


Figure 7: comparison across Experiments 1 and 3 of average d' scores for test conditions in which the face was old or new as a function of instructional condition. Dark bars represent the average d' scores for test items in the (F+V+) and (F+v-) conditions. Lighter bars represent the average d' scores for test items in the (f-V+) and (f-v-) conditions.

(e.g., when it provides information concerning the linguistic content of the speech event), since subjects were never told that faces would be important at later stages of the experiment.

The comparison between Experiments 1 and 3 confirms another of our hypotheses: overall, performance was always better when visual information was dynamic, compared with the scores from static visual displays. Dynamic visual information about the articulation of a word is simply more useful than static information in distinguishing old from new items, regardless of the encoding strategies subjects use for faces.

Finally, it is interesting to examine the effects of our instructional manipulation on the ability of subjects to use face information as a cue to recognition of speech events across experiments. In Experiment 1, static faces could *only* be used as retrieval cues for recognition when subjects were instructed to attend to faces in the experiment. When such explicit encoding instructions were not used, subjects were unable to utilize visual information in the process of distinguishing old from new items. We assume that the explicit instructions, then, cause subjects to consciously encode the visual information in a way which will be conducive to later, explicit recognition. As a result, an arbitrary link between the explicitly encoded face and the simultaneous speech event is formed in memory. In Experiment 3, dynamic visual information was only used as a retrieval cue for recognition when subjects were *not* alerted to the necessity for explicit recognition of faces later. *Explicitly* encoded dynamic faces could not be used as cues in the recognition of speech events, indicating that this different encoding strategy actually interferes with the natural, mandatory encoding of dynamic visual displays of the talker's articulation along with the speech signal.

The overall pattern of results from Experiment 3 indicates that dynamic visual information is encoded and stored in an obligatory fashion (i.e., without additional task demand) in memory representations for speech events. In addition, the results imply that tasks that require explicit encoding of dynamic face information may actually interfere with this automatic encoding, eliminating the potential utility of visual information in the recognition of words later. Although the potential use of *static* visual displays as retrieval cues can actually be *increased* by encouraging explicit encoding strategies, this increase in utility is not nearly as beneficial as that gained naturally by dynamic presentation, as shown by the relative proportions of variance accounted for by Face Context in Experiments 1 and 3.

Taken together, the findings from this experiment support the conclusion that dynamic information about talkers' articulation of a speech event is encoded integrally in memory with information about the acoustics of that speech event. In addition, the results indicate that explicit encoding strategies can interfere with this process. Given that a static face can only be used as an effective retrieval cue when the information is encoded explicitly, we conclude that studies that utilize static visual displays may not be taken as evidence for audiovisual integration in memory representations of speech events.

General Discussion

The present series of experiments was designed to examine the nature of multimodal speech representations in memory. In carrying out these experiments, we tested several hypotheses concerning the nature of the stimuli and the encoding tasks used to examine this important issue. Experiment 1 was designed as a control for the earlier Kato et al. (1995) experiments. Kato et al. (1995) used an explicit recognition memory procedure to determine whether static faces and arbitrarily paired voices were stored integrally in memory. Full sentences were used as the carriers of voice information. In contrast, the present set of experiments tested subjects ability to use implicit information about face-voice pairings by explicitly recognizing isolated words spoken in a mixed factorial design of studied and non-studied faces and voices.

It is important to emphasize here that the results obtained in all three experiments replicated the basic findings of earlier studies by Goldinger (1995) and Palmeri et al. (1991); words repeated in the same voice used during study were more easily recognized than words presented in a different voice. Additionally, Experiment 1 showed that manipulating the instructions given to subjects can alter their encoding strategies and thus can affect the way in which static faces can be used as retrieval cues in the recognition of spoken words. We found that explicitly instructing subjects that faces would be important later in the experiment encouraged them to encode stimuli in a way that fostered the use of static visual displays as effective retrieval cues in the later recognition of words. When no explicit encoding instructions were given, subjects were not able to use pictures of faces as additional retrieval cues for the recognition of words.

We conclude that under normal circumstances static pictures of faces are not encoded in memory along with speech events. Because the static faces were *arbitrarily* paired with the voice speaking a word during study and test trials, the information in the optical displays were not correlated with the underlying speech events to be encoded. As a result, a relationship between the arbitrarily associated visual and acoustical information was only encoded when subjects were alerted to the importance of the artificial, experimental relationship inherent to the stimuli. In other words, our explicit encoding instructions acted to artificially increase the potential utility of static faces during the experiment and, as such, subjects encoded them visual information in faces with respect to their relationship to simultaneously occurring speech events.

In contrast, Experiment 3 used *dynamic* movies of talkers articulating the word to be remembered. Since the acoustic specification of a speech event is lawfully tied to its articulation, the visual displays in these dynamic movies were informative with respect to the underlying speech event, because they contained optical information about articulation. As a result, subjects automatically encoded visual information in representations of speech events and were able to use this information later as effective retrieval cues in the recognition of spoken words.

However, when subjects were *explicitly instructed* that faces would be important later in the experiment, faces were not used effectively. The results suggest that the aspects of a face which can be most easily used for explicit recall are not necessarily the same aspects of a face which are potentially informative about an underlying speech event. Although the kinematics of a talker's speech articulators may provide a rich source of visual information concerning the speech being heard and seen, this dynamic information may not be useful in the explicit recall of that talker's face, and vice versa. As a result, biasing subjects to treat articulating faces as items for explicit recall may actually serve to reduce the potential of those articulating faces as information carriers by drawing attention away from articulatory aspects of the visual display and towards those aspects useful in the explicit recognition of faces.

A comparison of the results in Experiments 1 and 3 also revealed that, while static faces could be used as effective retrieval cues to the recognition of words, the gain in discriminability as a result of repeated face context was greater in Experiment 3, where faces were dynamic and provided complementary information about the utterance to be recognized. This finding indicates that while the potential relevance of a static face to the recognition of simultaneously presented speech events can be manipulated experimentally, the human perceptual and memory systems are more aptly designed to encode and exploit naturally occurring and lawful relationships between simultaneously occurring intermodal events (Fowler, 1986; Gaver, 1993).

Taken together, the present results contribute to the growing body of literature indicating that the processes of speech perception and spoken word recognition should not be viewed as exclusively auditory phenomena (see Bernstein, Demorest, and Tucker, in press, for a review of this literature). As noted before, the role of visual information in the process of speech perception was first described by the landmark study of Sumby and Pollack (1954), who showed that the intelligibility of speech in noise is greatly increased when subjects are allowed to see the talker articulate. Sumby and Pollack also showed that the relative contribution of visual presentation to overall intelligibility was independent of the speech-to-noise ratio under test, implying that the contribution of visual information is not simply *additive*, but *complimentary to* the auditory information presented. These initial conclusions were further supported by Erber (1969) who found that audiovisual information increased speech intelligibility even when speech is unintelligible when presented with auditory information alone. Both of these early studies suggest that acoustic and optical information can work in complimentary ways to support the perception of speech. The increase in speech intelligibility due to audiovisual presentation is equivalent to the gain in intelligibility afforded by an increase in speech to noise ratio of +15 dB (Erber, 1969; MacLeod and Summerfield, 1987; Middelweerd and Plomp, 1987; Rosenblum and Saldaña, 1996). This enormous gain in intelligibility clearly justifies further inquiry into the basis of multimodal speech perception and may provide important new theoretical insights into speech perception and spoken word recognition.

Further evidence of an integral role for vision in the perception of speech comes from the well-known McGurk effect (McGurk and MacDonald, 1976). With this illusion, McGurk and MacDonald showed that conflicting information in the auditory and visual modalities can significantly modify the perception of a speech sound and can even lead to a percept which is not specified by either modality alone (McGurk and MacDonald, 1976). For example, the presentation of an auditory [ba] with a visual [ga] led to the perception of [da] in 96% of the subjects tested. Because this multi-modal effect is extremely robust, subsequent studies have incorporated it into their designs in order to examine the integration of acoustic and optic information during the process of perception (Dekle et al., 1992; MacDonald and McGurk, 1978; MacLeod and Summerfield, 1987; Massaro, 1987; Massaro and Cohen, 1983; Massaro and Cohen, 1990; Munhall et al., 1996; Rosenblum and Saldaña, 1996; Rosenblum and Saldaña, 1992; Summerfield, 1984).

Taken together, these studies show that visual information, when available, is intrinsic to the process of speech perception. Indeed, a recent study by Bernstein et al. (in press) shows that visual information may even be *sufficient* for the process of speech perception, as demonstrated by the finding that some severely hearing impaired subjects can out-perform normal hearing subjects in identifying words presented with visual information only.

At the present time, several explanations have been proposed to explain the processes by which visual information affects the perception of speech. All of these explanations rest on assumptions concerning the nature of the information provided by the visual modality. For example, some explanations of the McGurk effect rest on the assumption that visual information and acoustical information allow for differing degrees of reliability in the perception of the speech sounds in question. Indeed, Summerfield (1987) presents evidence that the very speech segments which are most confusable when presented with solely auditory presentation are those which are least confusable when presented with solely visual information , and vice versa.

Massaro's (1987) formalization of this relationship, in terms of his Fuzzy Logical Model of Perception (FLMP), is based on the assumption that visual and acoustic information provide evidence with varying levels of reliability for the presence or absence of sub-phonemic features of the speech sound in question. Combining visually supported featural information with featural attributes which are supported by information from the auditory signal allows the perceptual system to derive a percept (e.g., Massaro, 1987; Massaro and Cohen, 1983). However, evidence has also been found which contradicts the proposal that information from disparate modalities is analyzed independently (Green and Kuhl, 1991). These results support a theory based on interactive processes in the perception of multimodal speech sounds.

Still, a theory of the analysis of sensory input cannot reveal much without addressing the fundamental problem of representation. As (Summerfield, 1987) points out:

> "Accounts of audio-visual speech perception must suggest how a knowledge of audio-visual structure is represented, and what information exists in the acoustical and optical streams to indicate that they should be interpreted together." (p.31)

A growing body of evidence suggests that the neural representation of speech must include information about the dynamic changes articulators make during the production of speech (Green and Gerdeman, 1995; Rosenblum and Saldaña, 1996). In a recent study, Rosenblum and Saldaña (1996) found that a point-light face display influenced the perception of speech only when it provided kinematic information about the talker's articulators, i.e., when it was moving. Static displays of point-light stimuli did not affect the perception of speech at all. Rosenblum and Saldaña (1996) argue that their findings contradict traditional theories of speech perception based on the perception and encoding of discrete featural cues, whether those cues be analyzed independently (e.g., Massaro, 1987) or interactively (e.g., McClelland and Elman, 1986). Indeed, it is hard to imagine how theories of perception based on units which have no extent temporally, like linguistic features or phonemes, can account for these findings.

In order to solve the problem of encoding temporal information in speech, it is worthwhile to consider the proposals of several investigators who have called for a drastic overhaul of the basic assumptions of most current theories of speech perception. One approach is the direct realist theory of speech perception advocated and elucidated by Fowler (1986). From this theoretical perspective, speech is perceived with respect to the event which produced it. An event is said to "structure" an informational medium (such as sound or light) in a lawful manner. Consequently, sound is viewed as a medium through which information about distal speech events may travel. The structure of this informational medium thus specifies the perception of a speech event. From the direct realist standpoint, then, it does not matter through which sensory modality information about a speech event comes, only that the information in that channel is related in some way to the event being perceived (Fowler, 1986).

In a paper presenting a new approach to auditory perception, Gaver (1993) argues for a conceptualization of the human auditory system designed to perceive the events and sources which cause sounds, not the sounds themselves. The structure of acoustic energy produced during a sound-making event is lawfully tied to the event which produced it and Gaver (1993) claims that the human auditory system may be structured in a way to exploit these relationships. While Gaver was more concerned with building a taxonomy of the events which can be perceived through sound than with explaining the process of speech perception, it is not hard to see how his ideas may be generalized easily to some of the long-standing problems in speech perception and spoken word recognition, since the acoustic energy which transmits speech is shaped and structured by the process of producing speech through the speech motor control system and the human vocal tract.

The present set of results suggests that the integrated encoding of multimodal sensory inputs is automatic and mandatory when the optic display available during a speech event is potentially informative about the underlying speech event. In Experiment 1 which used static faces, we found that explicitly instructing subjects about the potential usefulness of faces enabled them to make use of optical cues during later recognition of speech. However, for subjects who were not instructed in this way, static faces were shown to be very poor retrieval cues. Under normal circumstances, however, static faces are not encoded integrally with the speech signal, because there is no information in the acoustic signal that specifies a connection between the information in the two sensory modalities. In contrast, in Experiment 3, subjects who were not explicitly instructed about the potential usefulness of faces were able to utilize dynamic faces as cues in their recognition of speech events; explicitly instructed subjects could not. In this condition, perceivers were able to exploit and use the natural correlation between auditory and optical sources of information about a common underlying speech event.

Direct realism also takes a stand on the nature of information containing integrated sensory inputs; the information processed in the perception and encoding of speech must be *modality-neutral*, so that any information relating to an event can be used in perception, regardless of the modality through which it is obtained (Fowler & Rosenblum, 1991; Rosenblum and Saldaña, 1996; Summerfield, 1987). Whether this modality-neutral form is represented in articulatory terms or not, our results provide additional support for the hypothesis that the critical information is based on the time-varying properties of the speech signal (Rosenblum and Saldaña, 1996).

The findings obtained in these three experiments suggest that dynamic visual displays are automatically encoded in memory with auditory information in speech because they provide an additional source of information concerning the kinematic, time-varying properties of the underlying speech event. Dynamic visual information is lawfully tied to the form of acoustic information, and, as such, provides valuable cues about the source of the speech signal. This lawful bond between the disparate sensory modalities, to paraphrase Summerfield, (1987), serves as the information which indicates that the two input streams should be analyzed together because they are informative about the same underlying event. Explicit instructions to encode dynamic articulating faces may serve to divert the perceiver's attention away from those aspects of the visual display which are informative about the utterance and towards those aspects which can foster explicit recognition of faces in the memory task at hand.

The present set of findings, therefore, have several important implications for both memory and spoken word recognition. First, the evidence for multimodal encoding of speech presents serious conceptual problems for several well-known conceptualizations of human memory which operate on abstract, modality-specific inputs, such as Baddeley's phonological loop and visuospatial sketchpad (e.g., Baddeley, 1980). Given that all information entering long-term memory must first pass through working memory, it seems at best superfluous to assert that multimodal speech input, after being categorized using integrated representations, should once again be split into its component phonological and visuospatial parts for processing in working memory, only to be re-integrated later on for storage in long-term memory. An important new direction for future research on speech perception and spoken word recognition is to determine what the phonological loop "knows" about the speaker's face and how this kind of information is represented in working memory.

In addition, the evidence obtained here concerning different types of intermodal relationships (i.e., experimentally validated simultaneity conditions vs. natural, lawful and automatic interdependencies) and their relative value during recognition argues against any notion that information from disparate modalities is simply linked via some central executive function. In other words, it is hard to imagine within the context of Baddeley's tri-partite model of working memory, why a central executive, with the power to link representations of information from disparate sensory modalities, might do so differentially based on the relationship between them. Rather, it seems more parsimonious to posit a working memory system which operates on modality-neutral representations whose degree of integration is determined by task demands and exploitation of naturally occurring relationships. At the very least, the current conceptualization of the phonological loop must change from a system which operates on abstracted, discrete, symbolic linguistic units to a system that encodes and manipulates input which is informative about underlying speech events.

A second implication of the present findings concerns the structure of the mental lexicon. Landauer and Streeter (1973) have shown that there are important differences in the structure of similarity neighborhoods for high and low frequency words when similarity is measured by a metric that is based on the number of phonemes shared by two words. Luce (1986) and Luce and Pisoni (1998) found that these differences have ramifications for the perception and recognition of spoken words. Their findings suggest that the mental lexicon is functionally structured so as to reflect the phonemic similarity between words.

In a recent computational study that builds on the earlier work of Landauer and Streeter (1973), Auer and Bernstein (1997) analyzed the structure of the lexicon with all words recoded into strings of *visemes*. Their analysis revealed that a visually specified lexicon has a structure that "compliments" that of an acoustically specified one. That is, it seems as though spoken words which occupy dense *phonemically* specified similarity neighborhoods reside in relatively sparse viseme-coded neighborhoods, and vice versa (Auer and Bernstein, 1997).

If, as our study suggests, the neural representation of spoken words in lexical memory contains multimodal information, then the findings of Auer and Bernstein (1997) should have ramifications for the perception of words, just as the findings of Landauer and Streeter (1973) and Luce (1986) did for studies of spoken word recognition. In the same way that phonemic similarity has been shown to have effects on the perception of words, so too should similarity along visual dimensions. Thus, a serious reconceptualization of the underlying similarity space of spoken words in the lexicon is warranted, which takes into account both acoustic and optical features of spoken words.

The results of the present study demonstrate that speech perception is not necessarily confined to the realm of audition alone, but rather can operate in multiple sensory domains (Bernstein, 1998). One consequence of this view of speech is that it is no longer sufficient to base our theories of speech perception and spoken word recognition on auditory data alone, or to motivate those theories on acoustically-biased foundations. It is time now for a re-evaluation of the information we consider relevant to speech perception and of the processes we deem necessary for successful spoken word recognition.

References

Auer, E. T., & Bernstein, L. E. (1997). Speechreading and the structure of the lexicon: Computationally modeling the effects of reduced phonetic distinctiveness and lexical uniqueness. *Journal of the Acoustical Society of America*, **102**, 3704 - 3709.

Baddeley, A. D. (1997). *Working memory: Theory and Practice*. Boston: Allyn and Bacon.

Banks, W. P. (1970). Signal Detection Theory and Human Memory. *Psychological Bulletin*, **74**, 81 - 99.

Bernstein, L. E., Demorest, M. E., & Tucker, P. E. (submitted). Speech perception without hearing.

Campbell, R. (1994). Audiovisual speech: Where, what, when, how? *Current Psychology of Cognition*, **13**, 76 - 80.

Dekle, D. J., Fowler, C. A., & Funnel, M. G. (1992). Audiovisual integration in perception of real words. *Perception and Psychophysics*, **51**, 355 - 362.

Egan, J. P. (1958). *Recognition memory and the operating characteristic* (Technical Note Contract No. AF 19(604)-1962). Bloomington, IN: Indiana University Hearing and Communication Laboratory.

Erber, N. P. (1969). Interaction of audition and vision in the recognition of oral speech stimuli. *Journal of Speech and Hearing Research*, **12**, 423 - 425.

Fowler, C. A. (1986). An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics*, **14**, 3 - 28.

Fowler, C. A., & Rosenblum, L. D. (1991). Perception of the phonetic gesture. In Mattingly, I. G. & Studdert-Kennedy, M. (Eds.), *Modularity and the motor theory*.(33-59). Hillsdale, NJ: Erlbaum.

Gaver, W. W. (1993). What in the world do we hear?: An ecological approach to auditory event perception. *Ecological Psychology*, **5**, 1 - 29.

Goldinger, S. D. (1996). Words and Voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **22**, 1166-1183.

Green, K. P., & Gerdeman, A. (1995). Cross-modal discrepancies in coarticulation and the integration of speech information: the McGurk effect with mismatched vowels. *Journal of Experimental Psychology: Human Perception and Performance*, **21**, 1409 - 1426.

Green, K. P., & Kuhl, P. K. (1991). Integral Processing of Visual Place and Auditory Voicing Information During Phonetic Perception. *Journal of Experimental Psychology: Human Perception and Performance*, **17**, 278 - 288.

Kato, T., Kanzaki, R., Tohkura, Y. i., & Akamatsu, S. (1995). *Effects of other-mode context on face and voice memory* (Technical report of IEICE, PRU95-88, HIP95-15 (1995-07)). Kyoto, Japan: ATR Human Information Processing Research Laboratories.

Landauer, T. K., & Streeter, L. A. (1973). Structural differences between common and rare words: Failure of equivalence assumptions for theories of word recognition. *Journal of Verbal Learning and Behavior*, **12**, 119 - 131.

Legge, G. E., Grosmann, C., & Pieper, C. M. (1984). Learning Unfamiliar Faces. *Journal of Experimental Psychology: Learning, Memory and Cognition*, **10**, 298 - 303.

Lockhart, R. S., & Murdock, B. B., Jr. (1970). Memory and the Theory of Signal Detection. *Psychological Bulletin*, **74**, 100-109.

Luce, P. A. (1986). *Neighborhoods of words in the mental lexicon* (Research on Speech Perception Technical Report No. 6). Bloomington, IN: Speech Research Laboratory, Indiana University.

Luce, P. A., & Pisoni, D. B. (1998). Recognizing Spoken Words: The Neighborhood Activation Model. *Ear & Hearing*, **19**, 1 - 36.

MacDonald, J., & McGurk, H. (1978). Visual influences on speech perception processes. *Perception and Psychophysics*, **24**, 253 - 257.

MacLeod, A., & Summerfield, Q. (1987). Quantifying the contribution of vision to speech perception in noise. *British Journal of Audiology*, **21**, 131 - 141.

Massaro, D. W. (1987). *Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Massaro, D. W., & Cohen, M. M. (1983). Evaluation and Integration of Visual and Auditory Information in Speech Perception. *Journal of Experimental Psychology: Human Perception and Performance*, 9, 753 - 771.

Massaro, D. W., & Cohen, M. M. (1990). Perception of synthesized audible and visible speech. *Psychological Science*, **1**, 55 - 63.

Massaro, D. W., & Cohen, M. M. (1996). Perceiving speech from inverted faces. *Perception & Psychophysics*, **58**, 1047 - 1065.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, **18**, 1 - 86.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, **264**, 746-748.

Middelweerd, M. J., & Plomp, R. (1987). The effect of speechreading on the speech-reception threshold of sentences in noise. *Journal of the Acoustical Society of America*, **82**, 2145 - 2147.

Miller, G. A., Heise, G. A., & Lichten, W. (1951). The intelligibility of speech as a function of the context of the test materials. *Journal of Experimental Psychology*, **41**, 329 - 335.

Munhall, K. G., Gribble, P., Sacco, L., & Ward, M. (1996). Temporal constraints on the McGurk effect. *Perception & Psychophysics*, **58**, 351 - 362.

Palmeri, T. J., Goldinger, S. J., & Pisoni, D. B. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **19**, 309 - 328.

Parks, T. E. (1966). Signal-detectability Theory of Recognition-Memory Performance. *Psychological Review*, **73**, 44-58.

Pisoni, D. B., Saldana, H. M., & Sheffert, S. (1995). Multimodal encoding of speech in memory: a first report. In *Research on Spoken Language Processing Progress Report #20* (pp. 297-305). Bloomington, IN: Speech Research Laboratory, Indiana University.

Pollack, I. (1959). Identification of Elementary Auditory Displays and the Method of Recognition Memory. *The Journal of the Acoustical Society of America*, **31**, 1126-1128.

Rosenblum , L. D., & Saldana, H. M. (1992). Discrimination tests of visually influenced syllables. *Perception & Psychophysics*, **52**, 461 - 473.

Rosenblum, L. D., & Saldana, H. M. (1996). An audiovisual test of kinematic primitives for visual speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, **22**, 318 - 331.

Schacter, D. L., & Church, B. (1992). Auditory priming: Implicit and explicit memory for words and voices. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **18**, 915 - 930.

Sheffert, S., Lachs, L., & Hernandez, L. (1998). Hoosier Audiovisual Multitalker Database. In *Research on Spoken Language Processing Progress Report #21*. Bloomington, IN: Speech Research Laboratory, Indiana University.

Shepard, R. N. (1967). Recognition Memory for Words, Sentences, and Pictures. *Journal of Verbal Learning and Verbal Behavior*, **6**, 156-163.

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, **26**, 212-215.

Summerfield, A. Q., & McGrath, M. (1984). Detection and resolution of audio-visual incompatibility in the perception of vowels. *Quarterly Journal of Experimental Psychology*, **36A**, 51-74.

Summerfield, Q. (1987). Some preliminaries to a comprehensive account of audio-visual speech perception. In Aslin, R. N., Alberts, J. and Peterson, M. J. (Eds.), *The Development of Perception: Psychobiological Perspectives.*(219 - 255). New York: Academic Press.

Valentine, T. (1988). Upside-down faces: A review of the effects of inversion upon face recognition. *British Journal of Psychology*, **79**, 471 - 491.

Yin, R. K. (1970). Face recognition: A dissociable ability? *Neuropsychologia*, **23**, 395 - 402.

Some Observations on Working Memory Tasks and Issues in Cognitive Psychology[1]

Winston D. Goh[2]

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405

---

[1] This work supported in part by NIH-NIDCD Research Grant DC00111 to Indiana University Bloomington. The author thanks David Pisoni and Michael Vitevitch for helpful comments during the preparation of this article.

[2] Also Department of Social Work & Psychology, National University of Singapore.

Some Observations on Working Memory Tasks and Issues in Cognitive Psychology

Abstract. This paper reviews a variety of experimental paradigms that have been used in the study of working memory. Special emphasis is placed on issues that are relevant to language processing and comprehension. The first section describes the features and procedures of simple and complex span measures. The demand characteristics of these tasks and their implications for task performance are examined. The second section looks at experimental issues such as the use of closed and open sets of items in the memory span task, and the relevance of findings such as the phonological similarity effect and the word length effect on the nature of representation in working memory. The final section explores general theoretical issues that have emerged since the Baddeley and Hitch (1974) model of working memory. Directions for future research are discussed.

Introduction

The motivation for this review is to comment on some of the major issues that have evolved in the course of 25 years' research in the field of "working memory" (hereafter WM) since the seminal paper of Baddeley and Hitch (1974), which first postulated their now famous WM model. This model has three components: 1) a central executive, 2) the phonological loop, and 3) the visuo-spatial sketchpad. The central executive is an attentional system that controls and supervises the two subsidiary slave systems, the phonological loop and the visuo-spatial sketchpad. The visuo-spatial sketchpad is responsible for storing and manipulating visual images in WM, whereas the phonological loop handles speech-based information.

The phonological loop of WM has received the most attention from researchers and is currently the most developed component of the model (Baddeley, 1998). The loop consists of a passive phonological store and an articulatory rehearsal mechanism based on inner speech. WM traces held in the phonological store are assumed to fade after about 2 seconds unless they can be refreshed by subvocal rehearsal operations, which serve to maintain the integrity of the traces in the store. These rehearsal processes are also thought to be capable of converting written information into a phonological code and registering it in the store. This component of WM is thought to be important in language comprehension and may play role in learning new words (Baddeley, Gathercole, & Papagno, 1998).

Attempting to integrate *all* aspects of a quarter-century of work on WM would lead to a gargantuan tome, given that the topic is relevant to many fields within cognitive science. Rather, I have decided to concentrate only on those findings that might be relevant to language processing and comprehension. For the most part, the focus will be on selected papers from experimental psychology, supplemented by clinical findings where relevant. Language acquisition and developmental issues will be addressed at a later time.

This review is divided into three parts. I will begin by discussing the various WM tasks that have been employed by experimental psychologists and attempt a component analysis of these tasks. Part 2 will examine experimental issues and findings. Issues such as the nature of the items used in WM tasks, the nature of the WM code or trace, and the influence of long-term memory (LTM) on WM will be examined. I will then end with a discussion of general theoretical issues such as whether WM is separate from LTM and whether there are multiple WMs.

I. Working Memory Tasks

Simple Span Tasks

These tasks are called "simple" because they are not as cognitively demanding as the "complex" tasks, which typically involves the operation of a concurrent secondary task that places a cognitive load on participants. Simple span tasks tap the storage capacity of WM without any additional processing requirements that are characteristic of the complex span tasks, which will be discussed in the following section.

Immediate Memory Span. This is the traditional task used to measure a person's WM capacity. Digits, letters or words are presented in a list and the participant is required to report the items in sequential order. The size of the list increases until the participant is unable to recall all the items accurately or fails some other predetermined criterion. The various scoring methods will be discussed later. Cavanagh (1972) reported that the average immediate memory span for different items was 7.7 for digits, 6.4 for letters, 5.5 for words, and 3.4 for nonsense syllables. Forward digit span is, on the average, about 2 items higher than backward digit span (Lezak, 1995).

Word span does not appear to correlate highly with comprehension, as measured by the verbal Scholastic Aptitude Test (VSAT). Daneman and Carpenter (1980) reported a non-significant correlation of +.35. This poor relationship led researchers to develop more complex measures of memory spans (see complex spans below). However, more recent data by Engle, Nations and Cantor (1990) revealed higher correlations between simple word spans and VSAT, when the word frequency of the items in the task was controlled for. They found a correlation of +.63 between low frequency word spans and VSAT and a correlation of +.45 between high frequency word spans and VSAT. The implications of word frequency influences on memory span performance will be discussed in the section on long-term memory (LTM) influences.

La Pointe and Engle (1990) manipulated word length and found that immediate memory span for both short and long words correlated with VSAT with coefficients of +.37 and +.34 respectively. The magnitude of these correlations are comparable to the correlation between simple word span and VSAT obtained in the Daneman and Carpenter (1980) study, which did not manipulate word length.

Matching Span. The matching span task differs from immediate memory span tasks in that no overt reporting of the memory list is required. This makes the matching span task critical in determining receptive from expressive deficits for serial recall (Allport, 1984). Participants who only suffer from expressive deficits may perform poorly in immediate spans because of the inability to report the items, but would be able to perform normally on the matching span task. The task involves presenting a list of items (the memory set), followed by a target list, after which participants are required to respond "same" if the two lists are identical, or "different" if they are not. Typically, the lists are made different by reversing the order of two adjacent items (Allport, 1984). Because of the lack of overt reporting, matching span is usually longer than immediate memory span, presumably because overt reporting somehow interferes with WM traces in the latter task, resulting in a reduced span.

Missing Scan. Klapp, Marshburn, and Lester (1983) used this procedure to demonstrate that the capacity limit of immediate memory span is not a good measure of the limits of WM. Participants were presented with a list of 8 of the 9 non-zero digits and were required to report which digit was missing.

Presumably, participants had to compare the target list with an activated list of the 9 digits from 1 to 9 in WM to determine the missing number. Thus, the task taps the same resources as a simple digit span.

However, performance in the missing scan is not affected by rhythmic patterning (i.e., temporal grouping in which items are presented in groups of 3) or articulatory suppression. These two factors have been shown to strongly influence immediate memory span (Hitch, Burgess, Towse, & Culpin, 1996; Baddeley, Thomson, & Buchanan, 1975). Klapp et al. (1983) argued that the missing scan results suggest a WM component distinct from the component measured by ordered recall tasks such as immediate span. A major criticism of the procedure is that it has not been tested with letters and words. The use of a closed set of items (i.e., digits) might circumvent the use of WM resources since participants are always comparing the same memory set (the digits 1-9) with the target list. A procedure that varies the memory set might produce very different results.

Sternberg's Short-term Memory (STM) Scan. Although not strictly a measure of WM capacity, Sternberg's (1967) scanning task is a potentially useful procedure to determine the influence of item properties on the search process in WM. Previous research indicated that comparison rates are fastest for digits at 33 ms per item; and slowest for nonsense syllables at 73 ms, with letters, colours, words, geometric forms etc. in between (Cavanagh, 1972). One possible explanation for the variable comparison times is that the search rate depends on the amount of information encoded in the WM trace for each item. More complex items such as random forms and geometric shapes may be encoded and represented with more features than items such as digits. This would be consistent with the finding that multiple talker lists result in smaller memory spans than single talker lists (Saldana & Svec, 1995), because voice variation would result in the encoding of additional features. The Saldana and Svec findings, however, were limited to lists of letters and the extent to which it can be generalised to words, digits and other items would have to be investigated.

Complex Span Tasks

These tasks usually involve two concurrent processes: (1) a primary or criterion task which is similar to a simple span task in which a list of items is to be remembered, and (2) a concurrent secondary task which usually involves some complex manipulations or reasoning. The goal of these procedures is to measure the residual capacity of WM in performing the criterion task while putting the subject under a cognitive load with the concurrent task.

Reading/Listening/Speaking/Operations Span. Daneman and Carpenter (1980) developed the reading span task to obtain a WM measure that is more predictive of language comprehension. The task requires participants to read a set of sentences while remembering the last word of each sentence. After the end of the last sentence of the set, participants are required to perform a serial ordered recall of the last words. To ensure that participants are processing the sentences, they are either (1) required to read them aloud, (2) verify the truth of the sentence, or (3) verify the sensibility of the sentence. Reading span measures correlated .59 with VSAT and up to .90 with performance on fact and pronoun reference questions that tested comprehension of passages (Daneman & Carpenter, 1980).

The listening span test is similar, except that the sentences are presented auditorily. Similar correlations were obtained with measures on reading and listening comprehension of the passages. On reading comprehension, correlations with reading span measures ranged from +.74 to +.86, and from +.67 to +.72 with listening span measures. On listening comprehension, correlations with reading span measures ranged from +.42 to +.78, and from +.47 to +.85 with listening span measures. The reading and listening span measures are highly correlated in the region of +.80 and upwards (Daneman & Carpenter, 1980).

Another variant of this task is the speaking span (Daneman & Green, 1986), in which participants are given a list of words to remember, after which they are required to generate one sentence for each of the words in the correct serial order. Speaking span correlated +.57 with reading span.

Turner and Engle (1989) substituted mathematical operations for sentences as their secondary task to see if the effects of the concurrent task were due to linguistic processing. Operations span also predicted reading comprehension scores, producing correlations with quantitative SAT scores ranging between +.24 and +.33. Although simple digit span did not correlate with comprehension measures, it did correlate with the sentence and operation spans (ranging from +.18 to +.35; the complex spans inter-correlated with each other in the range of +.38 to +.58). Turner and Engle concluded that the nature of the secondary task is unimportant as long as it involved processing or reasoning of symbolic material. However, it should be pointed out that linguistic processing may also play a part in mathematical operations as verbal coding may be utilised (cf. Richardson, 1996).

Daneman and Merikle (1996) conducted a meta-analysis of studies using these complex spans and observed that the average correlational magnitude for comprehension ability was +.41 with sentence tasks and +.30 with operations tasks. In contrast, the correlations for simple tasks involving only the storage component were +.28 for verbal material and +.14 for digits. Thus, complex spans that require participants to engage in *both* processing *and* storage operations appear to be better predictors of comprehension than simple spans involving *only* storage operations.

Embedded Span. This paradigm was developed to determine if two successive lists of items interfered with the processing or storage of the other. Klapp et al. (1983) used this procedure to test if span memory and reasoning processes shared the same WM resources. Participants were given a primary task of memorising a list of letters, which was followed by an embedded task of memorising a smaller list of digits. They were then required to perform an immediate serial recall of the digits, followed by a delayed serial recall of the letters. It was hypothesised that if the two tasks shared the same WM capacity, there should be interference. The results confirmed this hypothesis, even when a consolidation period for rehearsal was given between the primary and embedded task. However, when the embedded task was a reasoning problem (e.g., 5 > 7, true or false?) or a search task using Sternberg's STM scan procedure, interference occurred only when no consolidation period was given for rehearsing the primary list. Klapp et al. (1983) interpreted this to mean that only the rehearsal processes interfered with reasoning processes, and that memory retention is independent of reasoning or search processes – provided that a consolidation process occurred before the secondary task. Therefore, retention of items in WM may not be using the same resources as other processes (although one could argue that retention uses the residual capacity of WM, which is not involved in the secondary tasks). This would mean that the full capacity of WM is not reached by the retention of items, as measured by simple span tasks.

Concurrent Span. This is a variant of the embedded span, with the two lists being presented simultaneously instead of consecutively. Only one of the lists is to be recalled, and the cue to determine which list to recall comes after the presentation of both lists. Thus participants must store 2 separate lists in WM. The paradigm allows the experimenter to manipulate the properties of each list to observe their effects on span performance.

Richardson (1984) found that imposing a second serial learning task reduces the phonological similarity effect in immediate serial recall. Learning a list of letters and words at the same time reduced the advantage of phonemically distinct words or letters over phonemically similar ones. This reduced advantage was more prominent when letters were to be recalled than when words were to be recalled. Richardson suggests that this result indicates a disruption of the efficiency of phonological coding by a concurrent load. However, it is unclear if a floor effect is confounding the results. Specifically, a concurrent memory load may compete for the same storage resources in the WM system. In this case, overall performance on both tasks may be impaired such that any advantage of phonological distinctiveness will be masked. Nevertheless, the concurrent span paradigm is potentially capable of revealing whether different aspects of WM are being utilised to process and store lists of differing items.

Component Analysis

A summary of the different WM tasks that I have discussed is provided in Table 1. Having described the nature of these tasks, some important questions can now be raised. What are the functional similarities among the various tasks, and what are the differences? To what extent can we infer the nature of WM processes and components from these tasks? The demand characteristics of most WM tasks can be divided into 3 broad categories: (1) encoding, (2) processing and, (3) output. Let us examine each in turn.

Table 1

Summary of WM Task Procedures

Simple Spans			
Name	*Task*	*Compare*	*Response*
Immediate Memory Span	$I_1, I_2, \ldots, I_n$		$I_1, I_2, \ldots, I_n$ $I_n, I_{n-1}, \ldots, I_1$
Matching Span	$I_1, I_2, \ldots, I_n$	$I_1, I_2, \ldots, I_n$ $I_2, I_1, \ldots, I_n$	Yes No
Missing Scan	$I_1, I_2, \ldots I_i, \ldots I_n$	$I_1, I_2, \ldots, I_n$	I_I
STM Scan	$I_1, I_2, \ldots I_i, \ldots I_n$	I_i X	Yes No
Complex Spans			
Name	*Task*	*Compare*	*Response*
Reading/Listening Span	$S_1 + I_1, S_2 + I_2, \ldots, S_n + I_n$		$I_1, I_2, \ldots, I_n$
Operations Span	$M_1 + I_1, M_2 + I_2, \ldots, M_n + I_n$		$I_1, I_2, \ldots, I_n$
Speaking Span	$I_1, I_2, \ldots, I_n$		$S_1 + I_1, S_2 + I_2, \ldots, S_n + I_n$
Embedded Span	$(I_1, I_2, \ldots, I_n)\ (J_1, J_2, \ldots, J_n)$		$(J_1, J_2, \ldots, J_n)\ (I_1, I_2, \ldots, I_n)$
Concurrent Span	$I_1, I_2, \ldots, I_n$ $J_1, J_2, \ldots, J_n$		$I_1, I_2, \ldots, I_n$
Key: I, J = items; S = sentence; M = math operation			

Encoding. Factors that affect encoding include (1) stimulus clarity, and (2) the number of stimulus attributes or amount of information per item.

Sternberg (1967) showed that stimulus clarity affected reaction time (RT) in his scanning STM task, such that degraded targets increased participants' RTs. This did not interact with the size of the memory set, suggesting that encoding and search processes are independent. Most of the other WM tasks do not vary stimulus clarity directly but there are methods used to disrupt encoding. For example, articulatory suppression (Baddeley et al., 1975), in which participants repeat an irrelevant word such as "the" during presentation and recall, is assumed to disrupt the encoding of the stimulus items into a phonological code, resulting in reduced spans. However, it is unclear if the locus of the articulatory suppression effect is due to a disruption of encoding or a disruption of rehearsal, because the latter is also affected by the irrelevant articulation required to perform this task.

The amount of information present in the stimulus appears to be more important than the actual memory set size. Saldana and Svec (1995) demonstrated that multiple talker lists resulted in shorter immediate memory spans than single talker lists, suggesting that information about the variability of the memory set is also encoded, rather than just the number of items. However, it appears that only salient information affects memory span, as multiple amplitude lists were no different from single amplitude lists. Indexical aspects of speech, such as voices, contain salient information in everyday language use, and are probably encoded holistically with the symbolic information of verbal material. This finding is consistent with research demonstrating that unattended speech, but not unattended noise, disrupts immediate memory span (Salame & Baddeley, 1987; 1989). Further research on the effects of the number of stimulus dimensions on memory span performance is needed in order to gain a better understanding of the factors that affect WM spans.

Processing and Retrieval. Factors affecting the processing and manipulation of material include (1) temporal order, (2) search and comparison, and (3) cognitive load. These factors are dependent on the nature of the tasks.

Temporal order refers to the sequence in which a list is processed and retrieved. In general, forward recall (in which participants are required to report the items in the order in which it was presented) has a longer span than backward recall. This is assumed to reflect the additional manipulation requirements of backward span. One possibility is that participants initially retrieve the list in a forward order and then read it off backwards. This additional requirement potentially uses more WM resources, resulting in a lower span. The immediate memory span task and most of the complex span tasks require ordered recall, with the latter exclusively in the forward direction.

For comparison tasks, participants are required to compare a target list or item with a presented list. It is assumed that a search of the WM store for the traces of the presented list occurs. These traces are then compared with the traces of the target items. Because these tasks do not require participants to overtly report a presented list, factors that affect processing, such as set size, are free from interference or disruption due to overt reporting. For example, memory spans as measured by a matching span task tend to be longer than the traditional digit span task. Comparison tasks allow the examination of search rates in WM by looking at reaction times as a function of set size. Capacity limitations can also be examined by looking at performance deterioration as the set size increases in matching span and missing scan tasks.

The complex span tasks, by definition, are the paradigms that impose an additional cognitive load by introducing a secondary task. This additional task presumably taxes the WM system if it requires the same resources as the primary task. Most of the primary tasks (which are effectively immediate memory span tasks) require ordered serial recall, although it is technically feasible to conduct matching or comparison tasks. For example, in a reading span task, one could present a list of words after the sentence set and ask participants if the list matches the last words from each sentence.

The reading and operations span measures were developed to predict language comprehension abilities rather than to distinguish among WM processes. The rationale for these measures was that people with better linguistic abilities would use less WM resources to process the secondary task, leaving more for the storage of items in the primary task. This would then be reflected in a higher residual WM span. Conversely, people with inefficient linguistic processing skills may devote more WM resources to process the secondary task, and thereby reducing their capacity for storing the primary task items. The embedded and concurrent span paradigms were developed to directly test whether WM processes involved in memorising a set of items interfered with another process that may or may not tap the same resources.

Output. The output required for the various WM tasks can be categorised as either (1) full report or (2) discrimination/identification. In the full report procedure, the entire list of items must be repeated. In the discrimination/identification procedure, either a binary response is made (i.e., yes/no or same/different) or a single item from the memory set is reported.

The full report procedure is used in the immediate span tasks and all of the complex span tasks. These tasks usually require ordered serial recall, although free recall may sometimes be used instead. The tasks involving comparisons and search, namely, the matching span, missing scan, and search paradigm, do not require an overt report of all the items in the memory set. Instead, these tasks use the discrimination/identification procedures which require only receptive ability and are presumably less susceptible to expressive impairment. Full report procedures, on the other hand, assume that *both* receptive and expressive abilities are intact.

Scoring Procedures

There are several ways to score performance in the various memory span tasks, Researchers may use one or more of these methods. Sometimes, researchers will report only one scoring method after stating that there were no differences among the different methods used.

1. Strict span score (Daneman & Carpenter, 1980) is the number of trials that are recalled perfectly on 2 out of 3 trials of a given set size or list length. If a subject obtained a perfect score on only 1 trial at the next higher set size, half credit (0.5 points) is given.

2. Absolute span score (La Pointe & Engle, 1990) is calculated by summing the total number of items in each perfectly recalled trial.

3. Total span score (La Pointe & Engle, 1990) is the total number of items recalled correctly, regardless of whether the trial was perfectly recalled or not.

4. Highest span score (Broadbent, 1971) is the largest set size or list length where all trials are perfectly recalled, rather than the size at which recall is 50% or more correct.

It should be pointed out that researchers differ in adhering strictly to the serial order of the items to determine whether a trial is perfect. Some will count words as being correctly recalled regardless of serial order (usually those employing complex span tasks, see Engle (1996)), whereas others will only consider trials in which temporal order is also maintained (which is the case for most simple span tasks). If temporal order is considered, some researchers will also report the number of items recalled in the correct serial position and the number of items in correct runs of 2 consecutive positions, in addition to the number of trials which are totally correct. Table 2 lists examples of the different scoring procedures.

II. Some Experimental Issues and Findings

Memory Span Items

We now turn to a discussion of item effects on memory span. That is, do the nature of the items in the memory set affect measures of memory span? Clearly they do. As discussed earlier, Cavanagh (1972) demonstrated that using digits as the memory set resulted in longer spans than letters or words. In addition, the average search time for each item in WM is inversely related to its immediate memory span. Items with longer spans have faster search rates. Thus, digits have the fastest search rate, whereas words have a slower search rate. Whether matching span times and missing scan times are also influenced by the nature of the items remain to be seen. Research using the matching span and missing scan paradigms has thus far reported only the list length scores but not search rate.

Table 2

Examples of Scoring Procedures

List Length	Trial	Presented List	Response
2	1	1 2	1 2
	2	3 4	3 4
	3	1 4	1 4
3	1	1 2 3	1 2 3
	2	2 3 4	2 3 4
	3	4 2 1	3 2 1 *
4	1	4 2 1 3	4 2 1 3
	2	1 2 3 4	1 3 2 4 *
	3	3 1 2 4	5 2 1 3 *

Strict Span Score: 3.5
Absolute Span Score: 6 + 6 + 4 = 16
Total Span Score: 6 + 8 + 6 = 20
Highest Span Score: 2

* incorrect items and/or order

Closed vs. Open Set Items. Closed set items refer to a group of items that form a distinct class whose membership is fixed. That is, no new items can be added to the set. Digits, letters, and function words are examples of closed sets. An open set refers to a group in which new members can be added. An example is the set of nouns. This distinction is potentially important for memory span studies. One of the problems of using closed sets or small vocabularies is that participants can combine partial information for a given item with their knowledge of the constraints within the set to develop guessing strategies

(Drewnoski & Murdock, 1980). For example, memory for the [I] sound in a digit span task might lead participants to infer that it was a 6 rather than 5 or 9. Another problem with using stimuli from a closed set is the increasing levels of proactive interference (PI) as items get repeated from trial to trial. Notice that these two factors predict opposite effects. A restricted range for guessing should lead to better span performance, whereas PI should impair performance.

A recent study by Coltheart (1993) showed that immediate memory span for words repeatedly sampled from a small pool was larger than words sampled without replacement from a large pool, provided that the words were phonologically dissimilar. There were no differences in performance when the words were phonologically confusable. This finding suggests that the restricted range facilitated performance rather than contributed to the effects of PI—the latter should have led to lower spans for repeated words relative to novel words. The fact that no difference was found for phonologically similar words lends further support for the restricted range argument, as there would be no guessing advantage of a restricted range if all the words sound similar.

It would be interesting to investigate this effect further with other WM tasks and with different memory set items. Most words used in such experiments do not form a naturally occurring closed set. Therefore, it may be useful to investigate this effect with words that form a true closed set such as numbers, letters or function words. One can also speculate that if more dimensions are added to a closed set list, it may reduce the advantage of the restricted range. Encoding an increasing number of attributes for each item may effectively increase the amount of information in WM, and as a result, increase the range rather than restrict it. This hypothesis is consistent with the letter span results obtained by Saldana and Svec (1995), which showed an advantage for single talker lists over multiple talker lists. If increasing the amount of information reduces the advantage of a restricted range, it may in turn increase the effects of PI, which could have been previously masked by the restricted range advantage.

It might be possible to examine this hypothesis with a release from PI paradigm (Wickens, Born, & Allen, 1963). When consecutive memory sets contain items from the same semantic category, recall performance drops because items from earlier sets interfere with the recall of items from the current set. However, if the subsequent memory set contains items from a different category, recall performance returns to initial performance levels.

It is assumed that participants are "released" from the effects of PI because the change in category is attended to. This means that any change, addition, or deletion of a stimulus dimension may be sufficient to cause a release from PI. With closed sets or repeated samples, the *addition* of another dimension will increase the range of the set by adding more information and thereby increasing the variance. This may have two consequences: 1) it may reduce the advantage of a restricted range, and 2) it may allow PI effects to set in. However, the *deletion* of a dimension may restrict the range even further and strengthen immunity from PI.

It is predicted that a switch from a list produced by a single-talker to a list produced by multiple talkers, while using a repeated sampling paradigm, would increase PI. This is because one is increasing the amount of information in WM by introducing more variation in the voices, and so restricted range advantages would be reduced. One would predict the opposite effect if a switch were made from a list produced by multiple talkers to a list produced by a single talker, while keeping the list items the same. In this case, the information in WM is reduced, allowing participants to benefit from the restricted range of a closed set.

La Pointe and Engle (1990) also explored the effects of repeated and non-repeated sampling. They found that the word length effect, which refers to the finding that memory span is inversely related to the length of the words (Baddeley et al., 1975), was not abolished by irrelevant articulation if the words were sampled from a large pool without replacement. This result casts doubt on the articulatory rehearsal explanation of the word length effect (Baddeley et al., 1975), which stated that concurrent articulation should abolish the word length effect regardless of repeated or non-repeated sampling. It is unclear why non-repeated sampling should be immune to articulatory suppression. One hypothesis is that novel presentations may encourage the use of more than one type of coding strategy that does not make use of the articulatory loop.

Representation in Working Memory

One of the major issues in WM research concerns the nature of the memory representation or code. In what form are items in WM stored? Although much research has investigated the storage and manipulation of visual images in WM, I will restrict this discussion to evidence related to linguistic material, given the goal of this review.

The Phonological Similarity Effect. Early research provided evidence for a phonological code for STM and a semantic code for LTM. Memory span for a list of phonologically similar words was found to be worse than the memory span for a list of phonologically dissimilar words or for a list of words which are semantically similar (Baddeley, 1966). However, for long-term learning and retention of words, semantic similarity enhances recall whereas acoustic similarity does not. The observed effect for similar sounding words in short-term recall is called the phonological similarity effect. This was the impetus for Baddeley to propose a phonological loop component that deals with verbal information in his WM model. Items entering this component are converted into a phonological trace. The discriminability of an item, therefore, depends on the extent to which its phonological features differ from the phonological features of other traces.

The Unattended Speech Effect. Further support for a phonological trace for WM comes from research that introduced irrelevant speech during the visual presentation of digits. Recall is equally disrupted by meaningful words, nonsense syllables, spoken digits, and words which contained some of the phonemes of the digits (e.g., *tun*, *woo* instead of *one*, *two*) (Salame & Baddeley, 1982). The overall results indicate that the unattended speech effect is caused by the phonological disruption of the WM trace, and not the semantic content. If semantic content were important, meaningful words and spoken digits should be more disruptive than the other two conditions. It was also shown that only verbal material will disrupt recall, as unattended noise did not have an effect (Salame & Baddeley; 1987, 1989).

The Word Length Effect. Another major finding in the WM literature is the word length effect. That is, memory span is inversely related to the length of the word. The locus of this effect is thought to be the duration of the word rather than the number of syllables. Lower spans are obtained when lists are comprised of words with a longer spoken duration such as *Friday* and *harpoon* than words such as *wicket* and *bishop* (Baddeley et al., 1975). This is consistent with the notion that phonological features, rather than semantic features, are the primary way that information is encoded in WM. Also, participants who use languages with a faster articulation rate tend to have longer digit spans. Digit spans tend to become shorter as one goes from Chinese, to English, to Malay, and to Welsh (Ellis & Hennelly, 1980; Naveh-Benjamin & Ayres, 1986; Hoosain & Salili, 1988; Elliott, 1992). Word length effects have also been demonstrated in both simple and complex span tasks (La Pointe & Engle, 1990). Baddeley (1998) attributes the word-length effect to the rate at which items can be rehearsed by the articulatory loop. The faster this can be done, the less the trace decays in the phonological store. Words with longer durations will reduce the number of traces that can be kept active, resulting in lower spans.

The explanation may turn out to be more complicated because lexical complexity and duration seem to have an influence on recall performance (Cowan, Wood, Nugent, & Treisman, 1997). Words which are more complex, as measured by the number of elements, such as disyllabic words, turn out to have an advantage over less complex monosyllabic words. Cowan et al. (1997) suggest that items defined by more information are more resistant to interference than less complex stimuli. Although information is lost through decay or interference, sufficient information is retained by the more complex items to cue the correct response. As in the section on closed vs. open set items, it appears that the amount of information attached to each item may be a critical factor for WM processes. Variables such as multiple talkers, multiple speaking rates, and multi-modal inputs may all have important influences on WM performance.

Another recent finding is that the word length effect appears to emerge only after extended trials in an immediate span task with repeated sampling. That is, there is no difference in performance between long and short words on initial trials (Nairne, Neath, & Serra, 1997). This suggests that the locus of the effect may involve proactive interference factors, and not simply a reflection of the duration of the WM trace.

Articulatory Suppression. The word length and phonological similarity effects can be abolished by articulatory suppression under certain conditions. Word length effects are generally negated by articulatory suppression (Baddeley et al., 1975). However, La Pointe and Engle (1990) found that this does not apply to words sampled without replacement. The phonological similarity effect is abolished only if the list presentation is done visually. This suggests that the articulatory loop is used to recode the items in the visual list to a phonological code. When the list is presented auditorily, articulatory suppression has no effect because the stimuli directly access the phonological store and do not need to be recoded (Baddeley, Logie, Nimmo-Smith, & Brereton, 1985).

However, the situation is not that simple. Bavelier and Potter (1992) found evidence for a phonological code that is impervious to articulatory suppression. When lists of words are visually presented at a very rapid rate of 100-117 ms per item, an effect known as repetition blindness causes the second occurrence of a word to be missed when the repeated word is presented in close proximity to the original presentation (Kanwisher, 1987). Orthographically similar words are usually affected (Kanwisher & Potter, 1990), but homophones are also susceptible to this (Bavelier and Potter, 1992). Articulatory suppression does not remove this effect, suggesting that recoding the visual stimulus into an auditory representation may involve some kind of very early phonological code that registers the item in WM. This code differs from the one used for recalling items from WM because it is not susceptible to articulatory suppression. More investigation into the existence of this code and the extent to which it is relevant to WM is needed.

Long-term Memory Influences

Lexical Effects. We are now heading into the realm of LTM influences on WM. Does the organisation and structural properties of words in LTM have an effect on WM processes? Engle, Nations, and Cantor (1990) demonstrated that word frequency influences memory span performance in both simple and complex span tasks. Memory spans were longer for lists comprised of high frequency words than lists comprised of low frequency words. However, the word frequency advantage was smaller in complex span tasks. A word frequency effect implicates LTM influences on WM because word frequency information can only be represented within LTM as a result of experience. Thus, immediate memory span measures for word lists may be affected by information stored in LTM.

Are there any items that can be used as a "pure" measure of WM that will not be contaminated by LTM knowledge? Some might argue for the digit span because the 10 digits are a closed set of items that are very familiar and very frequent. If this is true, one might argue for a function word span because these items are also a closed set that is very familiar and very frequent. However, performance on the digit span may be superior to performance on the function word span simply because people are more practiced at memorising a string of numbers in everyday life.

Other lexical properties of words influence recall in memory span tasks. Drewnowski and Murdock (1980) observed that intrusion errors tend to share auditory features with the omitted words. More detailed analyses revealed that these features tend to match the syllabic stress pattern and the stress placed on the vowel of the omitted word. Drewnowski and Murdock suggest that the vowel-to-vowel transitions might be a key feature of the auditory traces in STM. Evidence also suggests that phonotactic transition probabilities (Auer & Luce, submitted) and the neighbourhood structure of the words in the lexicon (Luce, Pisoni, & Goldinger, 1990) influence the speed and accuracy that words are accessed from LTM. Because WM is influenced by LTM, the effects that these properties of words in the lexicon have on word span tasks should be investigated.

A cautionary note is perhaps in order at this point. Lexical properties affect the ease of accessing the word from LTM. However, once the word is accessed and a representation stored in WM, it may be the case that LTM influences should no longer affect processing of the WM traces. Why should the neighbourhood structure of a word in one system (LTM) continue to have an effect when the word has been, in some sense, "transferred" to another system (WM)? The answer may be that the two systems are more intimately related than we realise. If lexical properties represented in LTM continue to exert an influence on WM, models that posit two functionally separate systems may no longer be viable.

Roles of Other Codes. There is also evidence for the use of codes other than phonological ones for representing items in WM. Wetherick (1975) explored the effect of using items from single versus multiple semantic categories in immediate memory spans. The results showed that recall performance was negatively related to the number of categories from which the items were selected. This result parallels the finding that multiple talker lists were detrimental to recall (Saldana & Svec, 1995). An effect of semantic categories suggests that participants were using pre-existing semantic relationships among words in LTM for recall. Wetherick postulated that if semantic information was coded directly rather than in a phonological form, there would be a clear advantage for single category lists because items from the same semantic category are structurally closer than items from different categories. On the other hand, the explanation offered by Saldana and Svec for the multiple talker effect is also a plausible alternative. More categories may imply more information to be encoded in WM, and this translates directly to the use of more WM resources per individual item, resulting in a smaller overall WM span.

Further support for a semantic component to WM comes from a series of neuropsychological case studies (Martin, 1993; Martin & Breedin, 1992; Martin & Romani, 1994; Martin, Shelton & Yaffee, 1994). These reports describe patients who have a severe deficit in immediate memory span, but have normal comprehension of speech and language. They are, however, unable to perform other tasks that require the maintenance of verbatim phonological material. This apparent dissociation suggests that the retention of items in the phonological store may involve a component of WM that is functionally separable from those involved in language and speech comprehension. For example, Martin, Shelton, and Yaffee (1994) proposed separate capacities for phonological and semantic retention after describing two patients who exhibited reversed patterns of deficits. The patient with the phonological retention deficit was impaired in sentence repetition but not sentence comprehension. However, the patient with the semantic retention deficit was impaired in sentence comprehension but not sentence repetition.

III. General Theoretical Issues

Is Working Memory Distinct from Long-term Memory?

There are three positions on this issue (Richardson, 1996). The original model of WM by Baddeley and Hitch (1974) and subsequent revisions take the position that WM is structurally separate from LTM. The WM system itself comprises distinct components which are themselves structurally separate from one another: (1) the phonological loop, (2) the visuospatial sketchpad, and (3) the central executive. Neuroimaging evidence suggests that there are portions of the brain that appear to be functional equivalents of these components (D'Esposito, Detre, Alsop, Shin, Atlas, & Grossman, 1995; Jonides, Smith, Koeppe, Awh, Minoshima, & Mintun, 1993; Kimberg & Farah, 1993). However, the neuroimaging studies do not address the issue of the separability of WM from LTM.

A second position argues that WM is merely the area of LTM that is currently active. Thus, it is not a structurally separate component of the human cognitive system. This account posits WM as the temporary raising of the activation levels of LTM structures. Anderson, Reder and Lebiere (1996) and Engle (1996) take this position.

A third position, advanced by Hasher and Zacks (1988), argues that the contents of WM are based on the activation of LTM representations. However, WM is still considered a distinct structural component of the information processing system. Interpretative processes within LTM activate the nodes and items relevant to the task at hand and these items are reflected in the WM component. This account is essentially a bridge between the other two.

Constraints on Working Memory

All three accounts agree that there are limited resources available for WM, but their different theoretical approaches conceptualise these capacity limitations in very different ways (cf. Richardson, 1996). For Baddeley's model, these limitations are due to the allocation of attentional resources by control mechanisms. This can be seen in performance during cognitively demanding tasks such as preload with concurrent processing. In this model, the word length effect is due to the rate of articulatory rehearsal. Articulatory rehearsal maintains information in the passive phonological store. Immediate memory span is believed to measure the storage capacity of this component.

For the other two approaches, which posit WM as the set of activated links in LTM, capacity limitations are due to the automatic spreading and decay of activation levels within LTM structures. These processes can be observed in the "fan" effects that are obtained for tasks that involve storing and retrieving facts. For these accounts, the current activation levels control the rate of processing. Those items that are more active will be processed faster than those items that are less active. Immediate memory span measures the sustained capacity of WM rather than the amount of information that is momentarily active. In this theory, the amount of information that is activated could be quite large. If the set of items exceed the capacity of WM, only those items with the highest activation levels will be reported.

Recently, Richardson (1996) pointed out that WM is likely to be constrained by both controlled attention and automatic activation processes. Which constraint is more important remains an empirical

218

issue. The importance of one constraint over another may be task dependent. However, for all accounts, the capacity of WM is not posited as a discrete quantity but an emergent property of the underlying processes. The *effective* WM capacity could be a result of the efficiency of the underlying processes and strategies that people employ in experimental tasks.

Inhibition and Interference

Inhibition refers to the processes and mechanisms that serve to prevent irrelevant information from gaining access to WM, thereby reducing the amount of available resources. Inhibitory mechanisms are a crucial feature of the Hasher and Zacks (1988) model of WM. They discovered that memory deficiencies found in older adults are not a result of reduced storage capacity but are due to a reduction in the efficiency of their inhibitory mechanisms. Hamm and Hasher (1992) found that older people maintain the different interpretations of ambiguous passages, even when the correct interpretation has already been implied. The work of Gernsbacher also implicates the use of inhibitory mechanisms (she uses the term *suppression*) in general cognitive processing. For example, Gernsbacher and Faust (1991) found that high comprehension participants dropped irrelevant meanings faster than low comprehension participants did. This, however, should be contrasted with other research by Just and Carpenter (1992), who found that high WM participants were able to retain irrelevant meanings of ambiguous phrases for a longer period of time than low WM participants. Engle (1996) pointed out that the crux of this issue is whether individual differences in WM capacity reflect inhibitory or activation processes. Gernsbacher and Faust (1995) argue that suppression is a result of active reduction in activation rather than passive decay or compensatory inhibition. This contrasts with the Just and Carpenter (1992) suggestion that lateral inhibition occurs when increased activation in certain nodes leads to the simultaneous decrease in activation of other nodes. Gernsbacher and Faust also argue that suppression varies with the task demands and the experimental context. This indicates that it is not an automatic and obligatory process, but may well be under strategic control.

It is also worth noting that the term *inhibition*, as used here, refers to active, directed, and effortful suppression (Bjork, 1989). Another use of the word inhibition refers to the effect of selecting one item over another because the activation level of one item is higher than the other. Bjork refers to this process as *blocking* rather than active inhibition – the item with a higher activation blocks the one with the lower activation, rather than inhibiting it. It remains to be seen whether both kinds of processes play a part in WM.

Inhibition can be contrasted with interference. Interference refers to properties between items that affect processing in WM. The phonological similarity effect is one example. Immediate memory span for items that are phonologically similar is much lower than for items that are dissimilar (Baddeley & Hitch, 1974). Semantic similarity among items also causes a reduction in immediate memory span relative to unrelated items, but not as profoundly as phonologically similar items. These results have been taken to indicate that the code of WM is phonological in nature, at least for the verbal component. However, if one takes the view that WM is the temporarily active portion of LTM and is not a distinct component of the cognitive system, it is unclear how these results would be accommodated. It is well known from priming studies that semantically related items could be considered to be stored "closer" together, so that activation spreads to these nodes faster than to unrelated nodes. Similarly, one could construe a distance metric for the phonological dimension. It is, therefore, unclear why properties of the items that facilitate activation in LTM should cause interference in WM. One possibility is that the locus of the effect is in the output mechanisms. Control processes that align the items for serial output, such as sub-vocal rehearsal, may be affected by certain item properties. Indeed, Klapp et al. (1983) argued that temporal grouping and articulatory suppression only affect the maintenance of serial order information. Tasks which do not require serial output, such as the missing scan task used in the Klapp et al. study, are unaffected by such factors.

The disparity between the influence of item properties on LTM and WM processes seem to indicate that at least some component of WM must be distinct from LTM. A study by Conway and Engle (1994) showed that although set size affected the time to scan for a target item (replicating the results of Sternberg's STM scanning experiments), it did not affect the amount of time required to retrieve the information from LTM. Although Engle takes the position that WM is just the active portion of LTM, it may be illuminating to see if retrieval time would be affected by phonological and semantic similarity among the items. According to past findings, item similarity should facilitate retrieval time from LTM but interfere with search in WM.

General vs. Multiple Working Memories

Is WM a single unitary component with a general-purpose pool of resources or is it a complex system with multiple domain-specific resource pools? If WM is a unitary component, resources must be flexibly allocated to the tasks at hand. However, if WM is a multi-component system, then the separate resources are independent of one another and cannot be reallocated to tasks outside of their specific domain.

Proponents of a general unitary WM system (e.g., Daneman & Carpenter, 1980; Just & Carpenter, 1992; Turner & Engle, 1989) point to the success of complex span tasks in predicting reading comprehension abilities. The conclusions of these studies generally appeal to the following rationale. The secondary processing tasks include a range of specific skills—mathematical ability, sentence processing etc.—which are not specific to reading. The primary storage task measures the residual capacity of the WM system after most of it has been allocated to the secondary task. As the predictive power of the measure does not change with the nature of the secondary processing task, it follows that all types of secondary task are tapping into the same general resource capacity, and whatever is leftover is used for storage. Thus, individuals with better general cognitive processing skills require less general resources for the secondary task, and have a greater capacity for the primary storage task.

Proponents of multiple domain-specific resources generally appeal to neuropsychological evidence showing functional dissociations and selective impairments of specific cognitive abilities. Separate components for semantic and phonological modules (Martin, Shelton, & Yaffee, 1994), arithmetical ability (Butterworth, Cipolotti, & Warrington, 1996), and visuo-spatial modules (cf. Baddeley, 1998) have been proposed. Richardson (1996) pointed out that the proponents of a unitary system are largely conceptualising WM at the level of Baddeley's central executive module, where the focus is on attentional systems that control processing of high level cognitive functions. Multiple component theorists are largely conceptualising the domain-specific resources at the level of Baddeley's subsidiary systems. The two approaches may not be entirely incompatible if there exists a component of WM that is responsible for executive functioning, but not short-term recall as measured by simple span tasks, and, conversely, a component of WM responsible for short-term recall but not executive functioning. Brain imaging studies have suggested that the frontal lobes may be the centre for executive processes, while other areas in the cortex appear to be responsible for the processing of verbal and spatial information. Future developments in this field may shed more light on the issue of unitary vs. domain-specific components in the WM system.

Conclusions and Future Directions

Several issues emerge from this selective review of WM research in experimental psychology. First, it appears that the amount of information within each item in a memory set influences memory span performance. It will be critical to uncover precisely which features of the stimulus are encoded in the WM trace, which aspects of the WM tasks these features will affect, and whether the saliency of these features depends on the level of processing required. For example, voice information affects immediate memory span procedures, but will it have the same effect on complex span tasks or those requiring scanning and identification of a target?

Second, inhibitory processes and interference effects must be seriously considered in accounts of WM. The effects of closed and open sets, repeated sampling and sampling without replacement may be contingent upon a complex interaction of information load, restricted range, and interference factors. WM performance may also be contingent upon individual differences in the efficiency of inhibitory mechanisms (Conway & Engle, 1994; Engle, 1996).

Third, the influence of LTM on WM deserves further investigation. No set of items used to measure WM span can truly be free from LTM influence. Even a memory span task using nonsense words will be affected by knowledge of the distribution of phonological forms in the language. Hence, it is important to investigate the extent to which LTM interacts with WM.

Finally, investigation of the role that WM plays in language processing must be conducted. It is clear that WM has a limited capacity; so for extended discourse, it is implausible that a person will be able to store verbatim the entire course of a language event. If language processing is dependent on this limited capacity, and if an increase in information load fills up this capacity, then the system can only handle this in one of two ways. One possibility is that the system could have evolved to favour the rapid extraction and inference of higher level linguistic units, while rapidly dropping the traces of lower level units once the higher units have been extracted. This method of operation serves to maintain a sustained WM capacity for incoming information and prevents overloading the system with too much information. Thus, the phonological components of a sentence may be rapidly dropped from WM once the words have been accessed, and so on, with only the semantic information of higher linguistic units kept in WM to await integration with incoming information. The other possibility is that linguistic information directly activates LTM representations. The LTM representations that are relevant to the current state of processing will then be active enough to "rise" to the conscious level of WM. It remains to be seen which account, or perhaps certain elements from both, will have greater explanatory power regarding the role of WM in language processing.

References

Allport, D.A. (1984). Auditory-verbal short-term memory and conduction aphasia. In H. Bouma & D.G. Bouwhuis (Eds.), *Attention and performance X: Control of language processes* (pp. 313-324). London: Erlbaum.

Anderson, J.R., Reder, L.M., & Lebiere, C. (1996). Working memory: Activation limitations on retrieval. *Cognitive Psychology*, **30**, 221-256.

Auer, E.T., & Luce, P.A. (submitted). Dynamic processing in spoken word recognition: The influence of paradigmatic and syntagmatic states. *Cognitive Psychology*.

Baddeley, A.D. (1966). Short-term memory for word sequences as a function of acoustic, semantic and formal similarity. *Quarterly Journal of Experimental Psychology*, **18**, 363-365.

Baddeley, A.D. (1998). *Human memory: Theory and practice* (revised ed.). Boston: Allyn & Bacon.

Baddeley, A.D., Gathercole, S., & Papagno, C. (1998). The phonological loop as a language learning device. *Psychological Review*, **105**, 158-173.

Baddeley, A.D., & Hitch, G. (1974). Working memory. In G.H. Bower (Ed.), *Recent advances in the psychology of learning and motivation* (Vol. VII, pp. 47-89). New York: Academic Press.

Baddeley, A., Logie, R., Nimmo-Smith, I., & Brereton, N. (1985). Components of fluent reading. *Journal of Memory and Language*, **24**, 119-131.

Baddeley, A.D., Thomson, N., & Buchanan, M. (1975). Word length and the structure of short-term memory. *Journal of Verbal Learning and Verbal Behavior*, **14**, 575-589.

Bavelier, D., & Potter, M.C. (1990). Visual and phonological codes in repetition blindness. *Journal of Experimental Psychology: Human Perception and Performance*, **18**, 134-147.

Bjork, R.A. (1989). Retrieval inhibition as an adaptive mechanism in human memory. In H.L. Roediger III & F.I.M. Craik (Eds.), *Varieties of memory and consciousness* (pp. 309-330). Hillsdale: Erlbaum.

Broadbent, D.E. (1971). The magic number seven after fifteen years. In R.A. Kennedy & A. Wilkes (Eds.), *Studies in long-term memory*. New York: Wiley.

Butterworth, B., Cipolotti, L., & Warrington, E.K. (1996). Short-term memory impairment and arithmetical ability. *Quarterly Journal of Experimental Psychology*, **49**, 251-262.

Cavanagh, J.P. (1972). Relation between the immediate memory span and the memory search rate. *Psychological Review*, **79**, 525-530.

Coltheart, V. (1993). Effects of phonological similarity and concurrent irrelevant articulation on short-term-memory recall of repeated and novel word lists. *Memory and Cognition*, **21**, 539-545.

Conway, A.R.A., & Engle, R.W. (1994). Working memory and retrieval: A resource-dependent inhibition model. *Journal of Experimental Psychology: General*, **123**, 354-373.

Cowan, N., Wood, N.L., Nugent, L.D., & Treisman, M. (1997). There are two word-length effects in verbal short-term memory: Opposed effects of duration and complexity. *Psychological Science*, **8**, 290-295.

Daneman, M., & Carpenter, P.A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, **19**, 450-466.

Daneman, M., & Green, I. (1986). Individual differences in comprehending and producing words in context. *Journal of Memory and Language*, **25**, 1-18.

Daneman, M., & Merikle, P. (1996). Working memory and language comprehension: A meta-analysis. *Psychonomic Bulletin and Review*, **3**, 422-433.

D'Esposito, M., Detre, J.A., Alsop, D.C., Shin, R.K., Atlas, S., & Grossman, M. (1995). The neural basis of the central executive system of working memory. *Nature*, **378**, 279-281.

Drewnowski, A., & Murdock, B.B., Jr. (1980). The role of auditory features in memory span for words. *Journal of Experimental Psychology: Human Learning and Memory*, **6**, 319-332.

Elliott, J.M. (1992). Forward digit span and articulation speed for Malay, English and two Chinese dialects. *Perceptual and Motor Skills*, **74**, 291-295.

Ellis, N.C., & Hennelly, R.A. (1980). A bilingual word-length effect: Implications for intelligence testing and the relative ease of mental calculation in Welsh and English. *British Journal of Psychology*, **71**, 43-51.

Engle, R.W. (1996). Working memory and retrieval: An inhibition-resource approach. In J.T.E. Richardson (Ed.), *Working memory and human cognition* (pp. 89-119). Oxford: Oxford University Press.

Engle, R.W., Nations, J.K., & Cantor, J. (1990). Is "working memory capacity" just another name for word knowledge? *Journal of Educational Psychology*, **82**, 799-804.

Gernsbacher, M.A., & Faust, M.E. (1991). The mechanism of suppression: A component of general comprehension skill. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **17**, 245-262.

Gernsbacher, M.A., & Faust, M.E. (1995). Skilled suppression. In F.N. Dempster & C.J. Brainerd (Eds.), *Interference and inhibition in cognition* (pp. 295-327). San Diego: Academic Press.

Hamm, V.P., & Hasher, L. (1992). Age and the availability of inferences. *Psychology and Aging*, **7**, 56-64.

Hasher, L., & Zacks, R.T. (1988). Working memory, comprehension, and aging: A review and a new view. In G.H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 22, pp. 193-225). San Diego: Academic Press.

Hitch, G.J., Burgess, N., Towse, J.N., & Culpin, V. (1996). Temporal grouping effects in immediate recall: A working memory analysis. *Quarterly Journal of Experimental Psychology*, **49**, 116-139.

Hoosain, R., & Salili, F. (1988). Language differences, working memory, and mathematical ability. In M.M. Gruneberg, P.E. Morris, & R.N. Sykes (Eds.), *Practical aspects of memory: Current research and issues, Vol. 2: Clinical and educational implications* (pp. 512-517). Chichester: Wiley.

BEST COPY AVAILABLE

Jonides, J., Smith, E.E., Koeppe, R.A., Awh, E., Minoshima, S., & Mintun, M.A. (1993). Spatial working memory in humans as revealed by PET. *Nature*, **363**, 623-625.

Just, M.A., & Carpenter, P.A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, **99**, 122-149.

Kanwisher, N.G. (1987). Repetition blindness: Type recognition without token individuation. *Cognition*, **27**, 117-143.

Kanwisher, N.G., & Potter, M.C. (1990). Repetition blindness: Levels of processing. *Journal of Experimental Psychology: Human Perception and Performance*, **16**, 30-47.

Kimberg, D.Y., & Farah, M.J. (1993). A unified account of cognitive impairments following frontal lobe damage: The role of working memory in complex, organised behavior. *Journal of Experimental Psychology: General*, **122**, 411-428.

Klapp, S.T., Marshburn, E.A., & Lester, P.T. (1983). Short-term memory does not involve the "working memory" of information processing: The demise of a common assumption. *Journal of Experimental Psychology: General*, **112**, 240-264.

La Pointe, L.B., & Engle, R.W. (1990). Simple and complex word spans as measures of working memory capacity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **16**, 1118-1133.

Lezak, M.D. (1995). *Neuropsychological assessment* (3rd ed.). Oxford: Oxford University Press.

Luce, P.A., Pisoni, D.B., & Goldinger, S.D. (1990). Similarity neighbourhoods for spoken words. In G.T.M. Altmann, (Ed.), *Cognitive models of speech processing: Psycholinguistic and computational perspectives* (pp. 142-147). Cambridge: MIT Press.

Martin, R.C. (1993). Short-term memory and sentence processing: Evidence from neuropsychology. *Memory and Cognition*, **21**, 176-183.

Martin, R.C., & Breedin, S.D. (1992). Dissociations between speech perception and phonological short-term memory deficits. *Cognitive Neuropsychology*, **9**, 509-534.

Martin, R.C., & Romani, C. (1994). Verbal working memory and sentence comprehension: A multiple components view. *Neuropsychology*, **8**, 506-523.

Martin, R.C., Shelton, J.R., & Yaffee, L.S. (1994). Language processing and working memory: Neuropsychological evidence for separate phonological and semantic capacities. *Journal of Memory and Language*, **33**, 83-111.

Nairne, J.S., Neath, I., & Serra, M. (1997). Proactive interference plays a role in the word-length effect. *Psychonomic Bulletin & Review*, **4**, 541-545.

Naveh-Benjamin, M., & Ayres, T.J. (1986). Digit span, reading rate, and linguistic relativity. *Quarterly Journal of Experimental Psychology*, **38**, 739-751.

Richardson, J.T.E. (1984). Developing the theory of working memory. *Memory and Cognition*, **12**, 71-83.

Richardson, J.T.E. (1996). Evolving issues in working memory. In J.T.E. Richardson (Ed.), *Working memory and human cognition* (pp. 120-154). Oxford: Oxford University Press.

Salame, P., & Baddeley, A.D. (1982). Disruption of short-term memory by unattended speech: Implications for the structure of working memory. *Journal of Verbal Learning and Verbal Behavior*, **21**, 150-164.

Salame, P., & Baddeley, A.D. (1987). Noise, unattended speech and short-term memory. *Ergonomics*, **30**, 1185-1193.

Salame, P., & Baddeley, A.D. (1989). Effects of background music on phonological short-term memory. *Quarterly Journal of Experimental Psychology*, **41A**, 107-122.

Saldana, H.M., & Svec, W.R. (1995, May). *The effects of talker-specific information on immediate memory span*. Paper presented at the 129th meeting of the Acoustical Society of America, Indianapolis, IN.

Sternberg, S. (1967). Two operations in character recognition: Some evidence from reaction time measurements. *Perception & Psychophysics*, **2**, 45-53.

Turner, M.L., & Engle, R.W. (1989). Is working memory task dependent? *Journal of Memory and Language*, **28**, 127-154.

Wetherick, N.E. (1975). The role of semantic information in short-term memory. *Journal of Verbal Learning and Verbal Behavior*, **14**, 471-480.

Wickens, D.D., Born, D.G., & Allen, C.K. (1963). Proactive inhibition and item similarity in short-term memory. *Journal of Verbal Learning and Verbal Behavior*, **2**, 440-445.

Familiarity, Similarity and Memory for Speech Events[1]

Sonya M. Sheffert and Richard M. Shiffrin[2]

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405

---

[1] This research was supported by NIH-NIDCD Research Grant DC00111, NIH-NIMH Research Grant MH12717 and NIH-NIDCD Training Grant DC00012 to Indiana University, Bloomington.

[2] Luther Dana Waterman Professor of Psychology, Indiana University, Bloomington.

Familiarity, Similarity and Memory for Speech Events

Abstract. The experiments described in the present report were designed to explore the "registration and learning" (Hintzman, Curran & Oppy, 1992) of instance-specific acoustic information. The purpose of this investigation was to determine if familiarity with a spoken word affects the acquisition of the perceptual details specifying a talker's voice and a word's plural form. It is generally assumed that the more we experience a word, the more complete our knowledge about the word will be. To test this hypothesis, the frequency of study presentation, the similarity between study and test items, and the instructions at retrieval were manipulated. Participants studied a list of words in which target items were repeated various numbers of times. In the test phase, participants heard a list of words that contained new items as well as old targets that were either the same as the studied item or were in a different voice, in a different plurality, or in a different voice *and* a different plurality. Subjects estimated the frequency of each word, with the additional caveat that they restrict their positive frequency judgments to items that were in the *same form* as the study item. The results revealed that presentation frequency improved listeners' knowledge that a word occurred, without improving their ability to discriminate it from a perceptually different word. This dissociation between knowing that a spoken word occurred and knowing perceptual details about the word suggests that instance-specific perceptual attributes of a spoken word are less likely to be encoded if the word has become very familiar over several repetitions. The collective import of this research is show that "registration without learning" occurs for auditorily presented words, for simple and complex stimulus dimensions, and across various experimental settings. The findings add to a growing literature on the factors that are important in determining what knowledge a subject encodes during spoken word recognition, and help to delimit theoretical interpretations of spoken language processing and memory.

Introduction

Among the most reliable findings in the literature on learning and memory is that familiarity facilitates recognition. Items that have become familiar over several prior presentations are recognized more accurately than new items or items presented only once. The research to be described here was motivated by an interest in discovering the changes in information processing that occur when an item is repeatedly presented to an observer. The aim was to determine if familiarity with a spoken word results in more efficient and complete acquisition of detailed perceptual information, or if instead, familiarity hinders the acquisition of perceptual information.

Memory for Repeated Events

A method researchers use to directly manipulate the amount of experience a subject has with an item is to vary the number of study repetitions the item receives. After the study list has been presented, a recognition memory or frequency judgment test can then be administered. Typically, well-known or often experienced items are recognized more accurately and given higher frequency estimates than items experienced once. An advantage the frequency judgment task has over a test of recognition memory is that the frequency judgment task provides information that is specific to the number of traces retrieved by a test item, rather than merely indicating the presence or absence on a trace. Moreover, the fact that judgments of frequency are quite accurate under a range of encoding and retrieval circumstances (Flexser & Bower, 1975; Greene, 1984; Harris, Begg & Mitterer, 1980; Hintzman & Block, 1971; Hockley, 1984; Naveh-Benjamin & Jonides, 1986; Rose & Rowe, 1976; Rowe, 1974), has led some theorists to assume that the information that underlies frequency judgments probably accrues naturally as an automatic consequence of event perception (Hasher & Zacks, 1974; 1984).

Registration Without Learning

It is generally assumed that the more we experience an item, the more complete our knowledge about the item will be. If it really is the case that "practice makes perfect," one would expect that our memory for highly familiar items should be very accurate. This intuitive idea was tested in a series of recent experiments by Hintzman, Curran & Oppy (1992). In one study, pictures were presented 1, 3, 8 or 15 times in a list. At test, the pictures were presented in the same orientation as at study or in a different (mirror reversed) orientation. Subjects estimated the frequency with which a study item occurred, with the additional requirement that positive frequency judgments should be given only to items that were in the same orientation. Hintzman et al. found that frequency judgments increased with presentation frequency. This indicates that subjects "registered" or catalogued the study repetitions in memory. However, given that an "old" judgment was made, the frequency ratings were similar for pictures tested in the same orientation or different orientation. In particular, Hintzman et al. found that there was no improvement in subjects' ability to distinguish same-orientation items from different-orientation items beyond the third repetition. This finding constitutes what has since come to be known as "registration without learning."

What is surprising about this result, according to the authors, is that someone can see a picture 15 times, "each time attending to it at least well enough to record the fact of repetition, without improving measurably in the ability to discriminate it from its mirror image" (Hintzman et al., 1992, p. 679). The same results have been found with other visual materials (singular and plural printed words). These effects are independent of listwise serial position, arising even when the first presentation of an item occurs near the end of the study list.

Hintzman & Curran (1995) have also shown that this bias against learning perceptual features on later repetitions is difficult to counteract. For example, simply telling subjects to attend to a particular stimulus dimension during study does not improve subjects' learning of the perceptual feature. Hintzman and Curran found that subjects must be made aware of the incompleteness of their knowledge before they show improvement in discrimination skill. This was accomplished by providing subjects with the opportunity to assess their own knowledge about the perceptual form of the word and, when incorrect, to focus additional attention on the relevant feature. As a consequence of this manipulation, their subjects showed a constant rate of learning over presentation frequencies.

Encoding of Repeated Items

Why would subjects fail to process a repetition as thoroughly as they do a first occurrence? The explanation proposed by Hintzman et al. (1992) is that repetition affects subjects' encoding processes. Specifically, familiarity leads to a redistribution of perceptual processing that changes the nature and extent of lower-level analyses (cf. DiGirolamo & Hintzman, 1997; Johnston & Hawley, 1994; Kraut, Smothergill & Farkas, 1981; Nickerson & Adams, 1979; Posner & Boies, 1971). According to this view, repeating an item activates previously stored memory traces of that item. The "priming" or activation that occurs as a result of repetition exerts a beneficial top-down influence on perceptual processing by decreasing the amount of sensory information needed to identify the item. This improves the speed and accuracy of item recognition.

However, the perceiver pays a price for this facilitation because processing of incoming sensory information is inhibited. The greater reliance on top-down conceptual information that accompanies familiarity decreases subjects' use of bottom-up data-driven information. Consequently, the stimulus attributes of a familiar item are less likely to be noticed. In order to draw attention away from conceptual features and towards perceptual features, aspects of the item's context must change appreciably across repetitions. In fact, Tulving and Kroll (1997) make the strong claim that a *necessary* condition for long-term storage of information is the perception of novelty. The redistribution of perceptual processing brought about by familiarity helps to optimizes behavior because it allows perceivers to capitalize on prior experiences, while also being able to readily detect and represent novel objects and events (Treisman, 1992).

The Retrieval of Repeated Items

Registration without learning has attracted considerable interest not only because it runs counter to our intuitions, but also because it runs counter to most current memory theories. The dissociation between knowing that a word occurred and knowing details about the perceptual form of the word suggests the use of two different retrieval processes (familiarity and recall). In contrast, most contemporary memory theorists (e.g., Gillund & Shiffrin, 1984; Hintzman, 1988; Humphreys, Bain & Pike, 1989; Metcalfe, 1982; Murdock, 1982; Ratcliff, 1978; Shiffrin & Steyvers, 1997) assume that recognition decisions (and, by extension, frequency judgments) can be modeled by a single process (familiarity) based on the match of a test item to the entire contents of memory.

In these memory models, memory is conceived of as a collection of episodic memory traces that represent each occurrence of a word and include information about the word's sensory form and context. Depending on the model, repetition can serve to strengthen the original presentation's trace (e.g., SAM [Raaijmaker & Shiffrin, 1981; Gillund & Shiffrin, 1984]), produce a new memory trace (e.g., Minerva 2 [Hintzman, 1988]) or both (e.g., REM [Shiffrin & Steyvers, 1997]). During retrieval, information about the test item's form and context combine into a single memory probe. Memory access is "global" in the sense that all the traces in memory are evaluated in parallel and contribute to the output of memory. The degree to which item and context features match is used as an index of the global activation or familiarity of the test item. Increases in familiarity result in an increase in the likelihood of identifying a test item as "old" and an increase in the magnitude of its frequency estimate. Likewise, similar distractor items (e.g., different orientation or different pluralization) that are falsely recognized will produce a response pattern akin to that of a less familiar target item.

Indeed, the registration without learning experiments confirmed this latter expectation (Hintzman et al., 1992). Hintzman's experiments showed that the judgments of frequency (JOFs) for similar distractor items increase proportionally to the frequency of target items. However, the experiments also revealed a large number of JOF=0 (i.e., "correct rejections"), a finding which is not expected if judgments are based on a single process. Hintzman infers that zero judgments occurred because sometimes subjects were able to explicitly recall enough information about the target item to allow them to rule out a similar test item. This "recall-to-reject" strategy requires a recall-like search component that supports the learning and recollection of item-specific information, in addition to the direct-access familiarity component responsible for registration.

Objectives of the Present Experiments

The broad theoretical implications of Hintzman's experiments highlight the importance of replicating and extending the "registration without learning" paradigm to a different domain and to new and more complex features. In addition, the literature on the effects of a perceptual match across study and test episodes on measures of explicit memory is based almost entirely on recognition memory tests (Richardson-Klavehn & Bjork, 1988). Therefore, it would be worthwhile to examine the retention of perceptual details using a different kind of explicit memory task.

The experiments described in the present report were designed to explore how experience with a spoken word affects the acquisition of instance-specific acoustic information. The first goal was to establish the generalizability of the "registration without learning" phenomena to the auditory domain. To this end, we developed an auditory version of visual plurality manipulation reported in Hintzman et al. (1992). The only major departure from Hintzman's design is that the study-to-test items vary by two dimensions (plurality, voice, or both), rather than one dimension. The experiments assessed the effects of word repetition on subjects' ability to make fine-grained distinctions among same-form and different-form items.

We also vary two types of features: voice and plurality. The motivation for comparing the effects of plural change with the effects of voice change was a suspicion that the lack of learning for plurality in Hintzman's experiments may have been related to the fact that plurality was a minimal or single feature. In contrast, There are reasons to expect that voice learning may be more robust than other stimulus dimensions and consequently may increase over word repetitions. One reason for this expectation is that the processing of talker information and phonetic information occur in a parallel-contingent fashion (Green, Tomiak, & Kuhl, 1997; Mullennix & Pisoni, 1990), whereas other sources of acoustic variation, such as speaking rate, amplitude (Bradlow, Nygaard & Pisoni, in press) and possibly plurality may not be processed integrally.

A second reason for this expectation is that a change in voice may be perceptually more salient than a change in plurality and therefore more likely to be noticed and encoded. The acoustic correlates that result from a change in voice are very complex, involving many acoustic-phonetic features, and affect not only the perception of the individual talker, but also the perception of the phonetic segments produced by the talker (Bricker & Pruzansky, 1976; Peterson & Barney, 1952). Moreover, there is ample evidence demonstrating that listeners encode details about a talker's voice in long term memory, and can explicitly recollect voice information during a word recognition test (Bradlow et al.; in press; Craik & Kirsner, 1974; Geiselman & Bellezza, 1976; Hintzman, Block & Inskeep, 1972; Palmeri, Goldinger, & Pisoni, 1993; Sheffert, in press; Sheffert & Fowler, 1995).

In contrast to voice, our plurality manipulation is acoustically very simple and unlikely to affect the perception of the preceding phonetic segments. Specifically, in the present study, pluralization of a singular word was accomplished by a simple acoustic transformation that involved splicing a [z] sound to the end of a word. The advantage of the splicing technique was that it eliminated the possibility of coarticulatory effects present in naturally produced plural tokens, and ensured that the only acoustic difference between singular and plural items was in the presence or absence of [z].

Using these stimulus materials, we manipulated the frequency of study presentation, the similarity between study and test items, and the instructions at retrieval. In Experiment 1, participants studied a long list of words that contained 24 target words produced by a male or a female talker. These words were either singular or plural. The targets were presented once or repeated five or ten times in the list. Participants listened to each word with the expectation that their memory would be tested after completing the study phase. In the test phase, participants heard a list of 48 successively presented words. Half were new items, and half were old targets that were either the same as studied items or were similar distractors. The degree of similarity between targets and distractors varied by one or two dimensions: Similar items were in a different voice, in a different plurality, or in different voice *and* in a different plurality from studied items.

In the test phase, subjects made frequency judgments based on *word token* information. That is, subjects estimated the frequency with which an item occurred at study, with the caveat that they restrict their positive JOF's to those items that were in the same form as the study item. For one group of subjects, "same form" meant same voice. They were told that test words were spoken in the same voice or in a different voice as at study, and that they should give "different voice" items a frequency judgment of zero. No mention was made of the plurality dimension. For another group of subjects, "same form" meant same plurality. They were told to give "different plurality" items a zero, and were not told about the plurality dimension. Subjects were only told about one of the two dimensions. This was done in order to determine if a change in the irrelevant dimension would influence frequency judgments or same-form/different-form discriminations[3]. Together, our design allows us to replicate and extend Hintzman's experiments to the auditory modality and to determine the extent to which feature complexity, feature integrality and test instructions moderate the "registration without learning" phenomena.

Experiment 1

Method

Participants

The participants were 48 Indiana University students who volunteered for the experiment in exchange for course credit. Half received the voice instructions and half received the plural instructions. All were native speakers of English and reported no history of any speech or hearing disorders at the time of testing. The participants were tested in groups of four or fewer. The data from four voice and two plural subjects were eliminated owing to a failure to follow instructions.

Apparatus

Presentation of the stimulus materials and collection of the response data were carried out for each participant on an IBM-compatible personal computer. The words were presented binaurally to subjects at 75 dB SPL over matched and calibrated stereophonic headphones (BeyerDynamic DT-100).

Stimulus Materials

The target stimuli were 24 monosyllabic singular words produced by a male speaker and a female speaker. The words were selected from the Indiana University Multi-talker Speech Database (see Torretta, 1996, for a detailed description). The audio recordings were obtained by asking the speakers to read each item in citation form. The recordings were made in a sound-proof booth, and then digitized at a 20 kHz sampling rate and equated for peak amplitude.

---

[3] An obvious third condition is one in which subjects restrict their positive frequency judgments to those items that were in the same voice and the same plurality as the study item. This experiment was conducted. We found that under these conditions, subjects favored a strategy whereby they judged most target items to be in a different form, even those that were in exactly the same voice and plurality. This response bias produced an inordinate number of zero judgments, and effectively rendered the data uninterpretable. Consequently, the complete results from this experiment will not be reported.

Plural forms of the singular target words were created by splicing a sample of word-final frication onto a singular word. For example, /z/ was spliced onto the end of the word *dog* to produce *dogs*. To ensure that the plurality of target items was perceptually salient, a pretest was conducted which assessed the intelligibility of the singular and plural forms from each talker. A separate group of five listeners (two speech scientists and three undergraduates) identified each word. Overall intelligibility, as defined by the percentage of correct word identification, was quite high (95%), and, importantly, plural items were unanimously heard as such.

The study list consisted of 24 distinct target items, with eight assigned to each of the presentation frequencies: 1, 5, and 10. Half of the items were presented in a female voice and half in a male voice. Similarly, half of the items were singular and half were plural. The form of an item was consistent across all presentations. Four different study lists were constructed in which each of the 24 target items was rotated through a different, randomly selected condition. An additional 77 items were randomly selected as filler items, and were used to decrease the possibility of spacing effects on explicit memory. The filler items were produced in the same two voices as the target items and included both singular and plural items. Each filler item occurred only once in the study list. Subjects were randomly assigned to study lists, and the order of presentation was random for each subject. Each study list was 205 items long.

The test phase consisted of 24 target items and 24 new items. Target items were of four types: Six target items were repeated in the same voice (designated as "same" items), and were the exact tokens used in the study phase. Six target items were repeated in a different voice ("different-voice"); six were repeated in a different plurality ("different-plural"); and six were repeated in a different voice *and* in a different plurality ("both-different"). An additional 24 items were new fillers. Subjects received one of four different test lists, depending on the list they studied. The order of words was individually randomized for each subject.

Procedure

There were two phases in the experiment: study and test. The study list consisted of 205 successive word presentations. Subjects were told that some of the words would be presented more than once in the list. They were also told to try to remember each occurrence of the word because they would receive a memory test at the end of the study session which would require them to estimate the number a times each word was presented. After listening to each word (which typically took approximately 1.5-2 sec), participants pressed a key to signal the computer to begin playing the next word. The interval between this key press and the onset of a new trial was 1 s. The study phase was followed by a short (2-2.5 minutes) retention interval, followed immediately by the test phase.

The test list consisted of 48 successive word presentations. The order of presentation was random for each subject. Participants in the voice instruction group were told that they would hear a list of words, some of which were present on the study list, and some of which were new. They were told that their task was to estimate the number a times a word was presented during the study phase. Following Hintzman et al. (1992), subjects were explicitly instructed to pay close attention to whether the word was in the same voice as at study, and were told to only type the number of times they heard the word exactly as heard on the test list. Subjects were provided with the following example:

> *Your task is to estimate the number a times a word was presented during the study phase. Only type the number of times you heard the word exactly as heard on the test list. Pay close attention to whether the word was spoken by a male or female voice For example, in the study list you could have heard the word "cat" spoken by the female twice, but never "cat" spoken by a male The correct answer depends on whether "cat" (the female version) is presented to you in the test. If "cat" (the original female version) is in the test, you would be correct if you answered "two". If "cat" (the male version) is on the test, you should answer "zero" because that exact word was not presented. Finally, the form of a word was consistent across word repetitions.*

The complement to these instructions were given to subjects in the plural instruction group. The instructions focused attention on the plurality of the item, rather than the voice. Subjects in both groups were also told that no word was ever presented more than 15 times, so all their frequency judgments should be between 0 and 15. Each time a word was presented, the subject had to type their response using the numeric keys. When they completed their frequency judgment, the subject would press the return key and the next trial would began 1 second later. The entire experiment lasted approximately 30 minutes, after which the subjects were debriefed.

Results and Discussion

Frequency Distributions

The purpose of this measure is to determine if the shape of the response distribution differed for items that were identical to the studied words as compared to items that were similar to the studied words. If a single process (familiarity) underlies frequency judgments, the response distributions for similar items should resemble scaled down versions of the response distributions of same items.

Figures 1a and 1b displays the frequency judgment distributions for each item type from frequency = 10. These patterns are representative of all three frequency conditions. Figure 1a represents the voice instruction group. In that figure, the top two panels (same and different-plural) show a unimodal response distribution, with the majority of responses between 6 and 12. The bottom two panels (different-voice and both-different) show a bimodal response distribution, with the first mode at 0, and the second mode between 6 and 12. The large number of 0 responses reflect subjects' ability to reject similar distractors, and indicate that some detailed information about the physical form of words was encoded during the study phase. The remaining nonzero JOFs are "false recognitions" and reflect subjects' failure to discriminate same items from similar items.

Insert Figure 1a about here

A complementary pattern was found in the plural instruction group (shown in Figure 1b). Same and different-voice items produced unimodal distributions, and different-plural and both-different items produced a bimodal distribution.

Insert Figure 1b about here

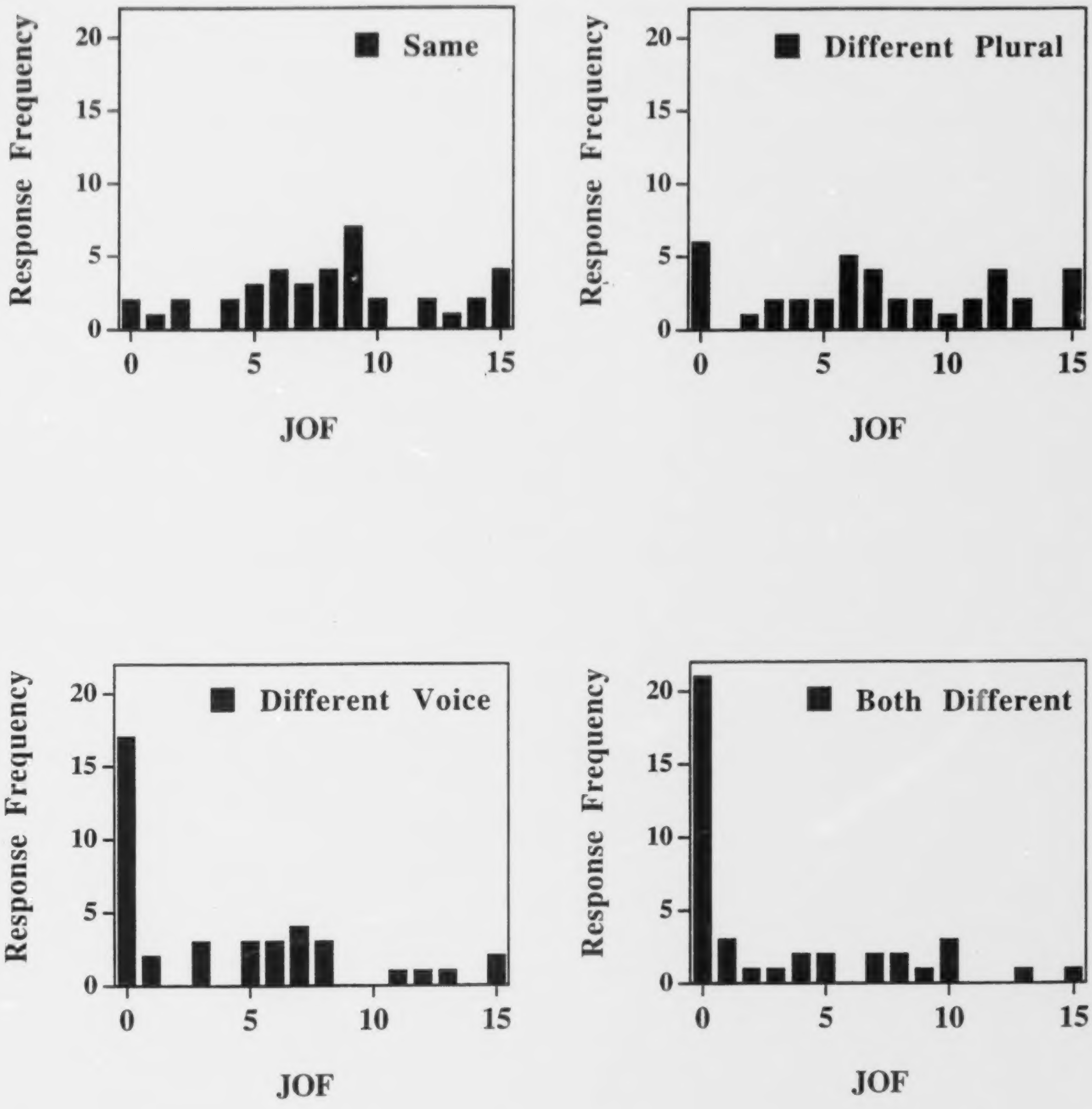


Figure 1a. Histograms of the frequency judgment distributions for subjects in the voice instructions condition. The figure displays the number of responses for each item type at presentation frequency = 10.

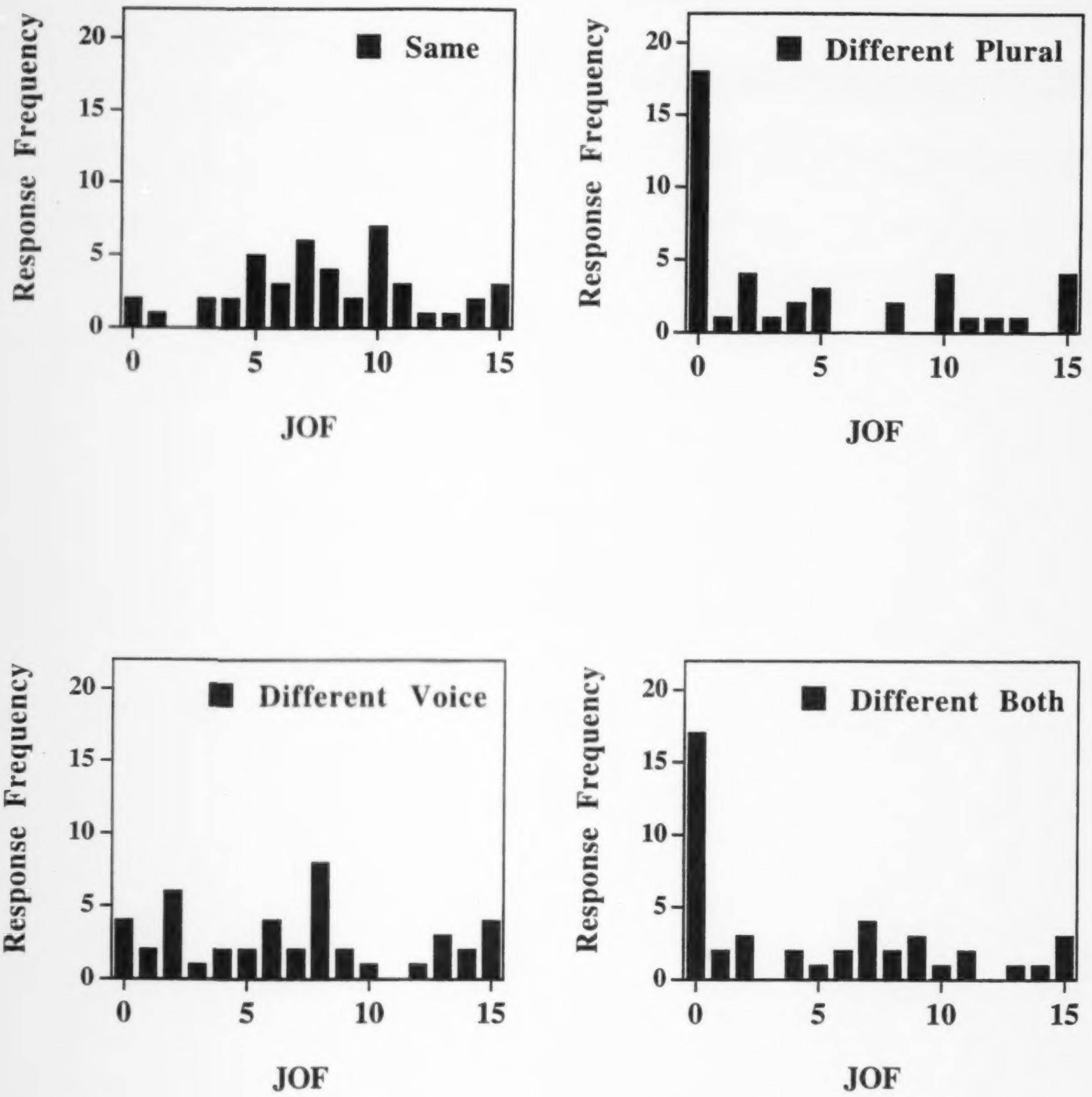


Figure 1b. Histograms of the frequency judgment distributions for subjects in the plural instructions condition. The figure displays the number of responses for each item type at presentation frequency = 10.

The response distributions also reveal an effect of instructions. Figure 1a shows that subjects were largely unaffected by study-to-test changes in the plurality under conditions that required them to discriminate items based on a voice match. Compatibly, Figure 1b shows that subjects were unaffected by changes in voice under conditions that required them to discriminate words based on a plurality match. Thus, although subjects were not explicitly told to ignore the irrelevant stimulus dimension when making their same form-different form discriminations, they nevertheless did so. This pattern was found in several of the following measures.

Mean Frequency Judgments

This measure provides information about subjects' sensitivity to variation in presentation frequency and whether JOFs overestimate or underestimate the actual presentation frequency. Because the response distributions of the different-form items were not normally distributed, only the mean JOF's from the same items were analyzed. Mean JOFs for same items are displayed as a function of frequency and instruction condition in Figure 2. The positive slope of the lines indicates that subjects in both groups were sensitive to presentation frequency. Mean JOFs across the three frequency levels (collapsed over instructional group) were 1.25, 4.2 and 8.1. An ANOVA on the frequency estimates from each test group revealed a highly significant effect of frequency for the voice group [$F(2, 38) = 39.45$, $p < .0001$] and for the plural group [$F(2, 22) = 82.56$, $p < .0001$]. A separate analysis comparing frequency estimates across the two groups showed that there was no difference across instructional conditions (this holds true for all subsequent analyses).

Insert Figure 2 about here

Nonzero JOFs

This measure assessed whether items that were falsely recognized (given a JOF of greater than 1, rather than 0) were perceived as less familiar, and consequently given a lower JOF than items that were correctly recognized. The analysis excluded all JOF = 0, and analyzed only the nonzero frequency judgments.

Figure 3 displays the mean nonzero JOFs as a function of test condition, presentation frequency and item type. The figure shows that JOFs increased as actual frequency increased, and that overall performance levels were similar across the two test conditions (the mean nonzero frequency judgment was 4.6 for the voice group and 4.7 for the plural group). The figure also shows differences among the four item types at frequency = 10, but only in the voice instruction group.

Insert Figure 3 about here

An ANOVA was conducted on the nonzero frequency judgments at each frequency level, separately for each test condition. Each analysis included the within-subject factor of item type. Subjects who failed to contribute at least one nonzero judgment for each item type at each frequency were excluded from the analysis. This was necessary to prevent subject selection artifacts from contaminating the inferential statistics (see Hintzman and Curran, 1994).

236

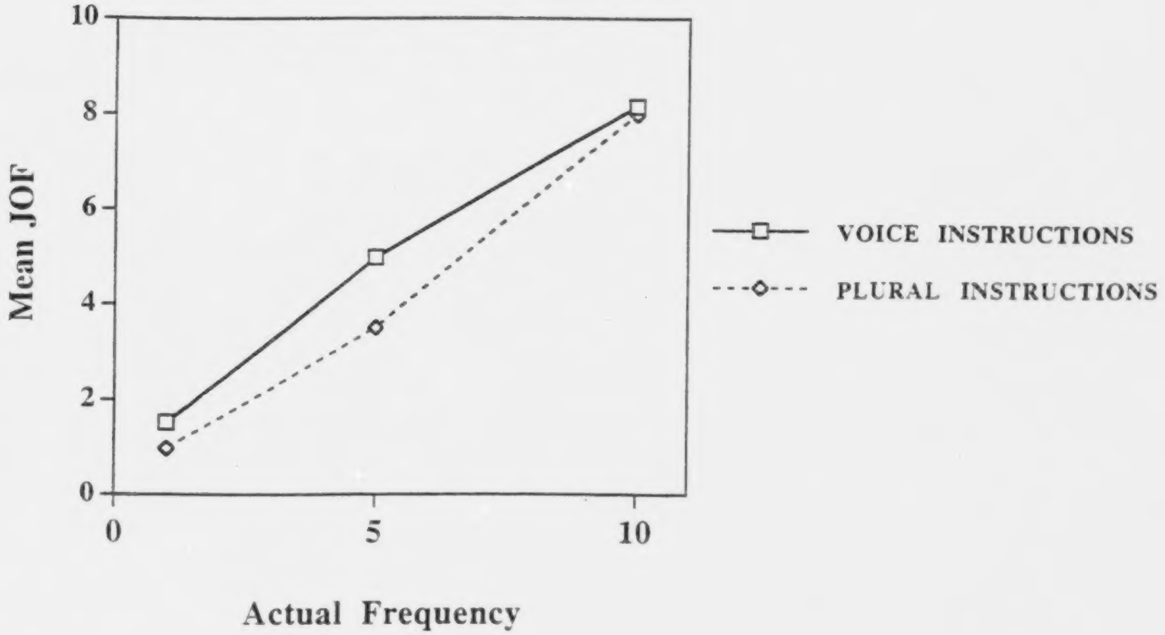
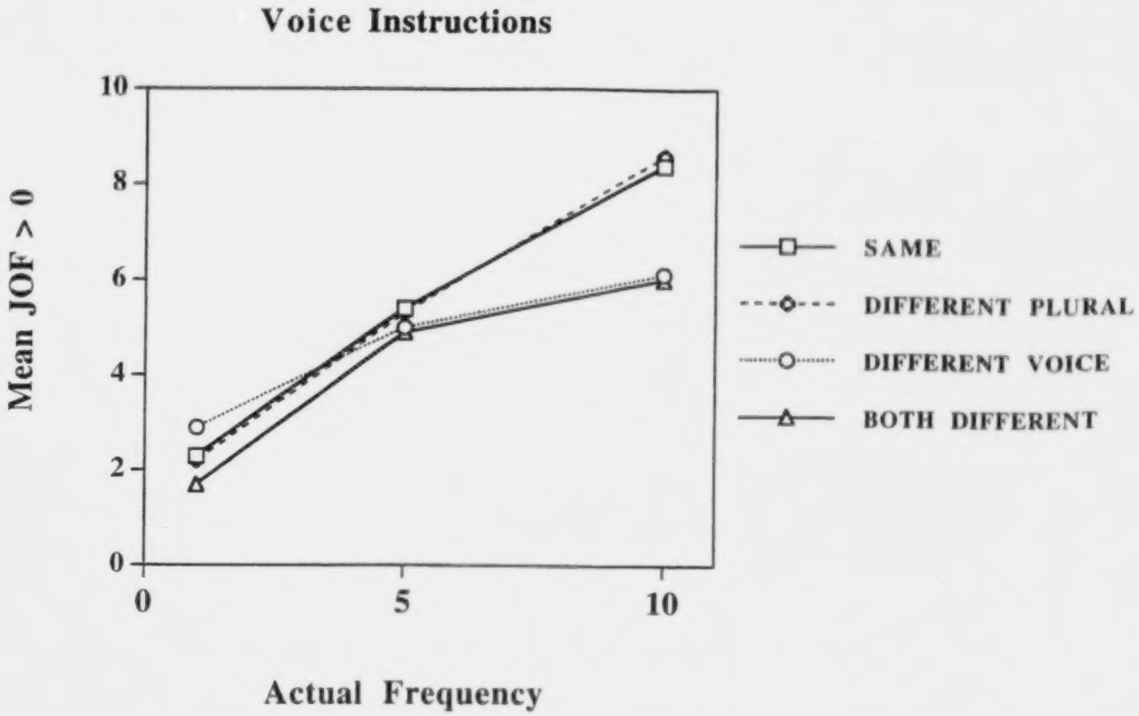


Figure 2. Mean judgments of frequency (JOFs) are displayed as a function of actual frequency and test instruction group, collapsed over test item type.



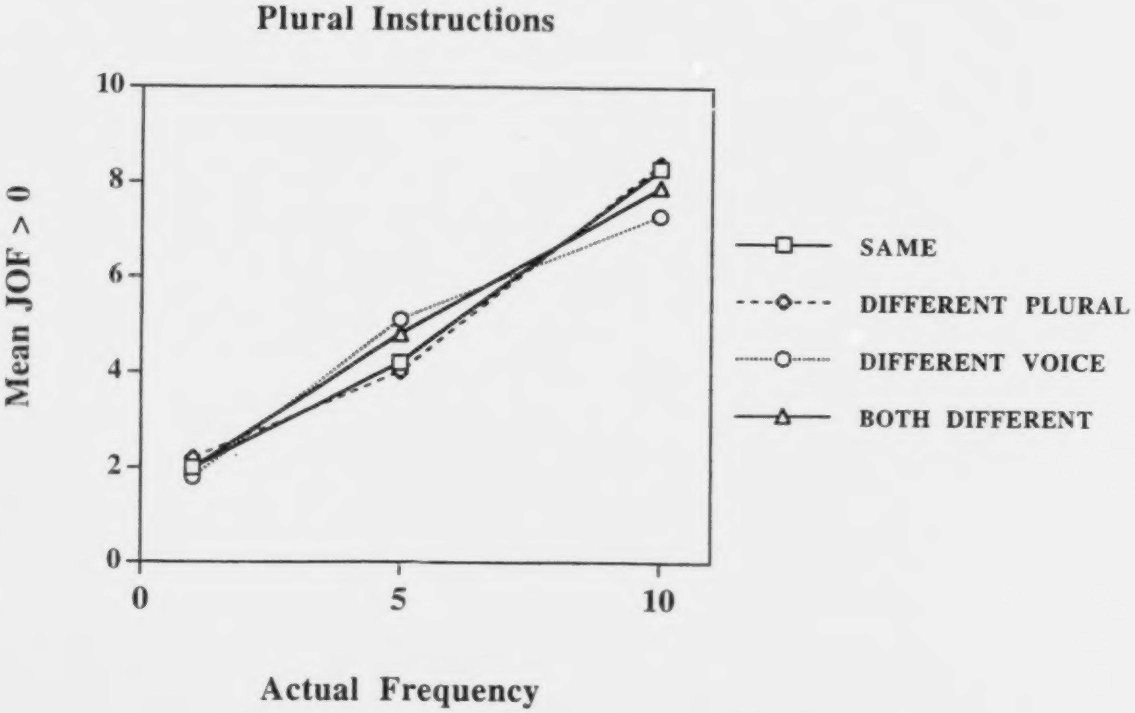


Figure 3. Mean judgments of frequency (JOFs) greater than zero are displayed as a function of presentation frequency and test item. The upper panel displays frequency judgments for subjects in the voice instructions condition. The lower panel displays frequency judgments for subjects in the plural instructions condition.

For subjects in the voice instruction condition, item type was significant at frequency = 10 [F(3, 42) = 3.06, p < .04], reflecting the fact that the frequency judgments for same items were higher than frequency judgments for similar items. Specifically, nonzero frequency judgments were higher for same items than for different-voice items [t(17) = 2.41, p < .03], and were higher for different-voice items than for different-plural items [t(17) = 2.88, p < .01]. Thus, study-to-test changes in voice led to lower JOF's, presumably because the words were less familiar than same-voice items. This perceptual match effect is similar to the pattern obtained in Hintzman et al. (1992; 1995), except that his data reveal differences between same and similar items at lower frequencies.

For the plural instruction group, the similarity between study and test items was not reliable at any frequency, which indicates that same and similar items were perceived as equally familiar. One reason for this difference across instructional conditions may be that plural information is less likely to be encoded in memory than voice information. This hypothesis is addressed in a subsequent word recognition measure and directly tested in Experiment 2.

Alternatively, subject variability may be the reason for the null effect of item similarity in the plural group, as well as the weak effect at lower frequencies in the voice group. Examination of the individual subject data reveals extremely variable performance across subjects. Consequently, it may be necessary to run as many subjects as Hintzman (typically 70 subjects per condition) in order to replicate his findings.

Word Recognition

In this analysis, frequency judgments were collapsed into a binary recognition measure. The recognition measure was obtained by classifying all words given a judgment greater than 0 as "old" and those given a judgment of 0 as "new". Figure 4 displays the proportion of JOFs that were greater than 1 (JOF > 0) as a function of presentation frequency and item type. The top panel shows the data from the voice instruction condition and the lower panel shows the data from the plural instruction condition. The corresponding *d'* values for both conditions are provided in Table 1.

Insert Figure 4 about here

For subjects in the voice instruction condition, if a test word was in the same voice as at study (which would be the case for same and different-plural words), subjects typically followed the instructions and gave it a frequency judgment that was greater than zero. As the figure shows, there was a large proportion of JOF > 0 for same voice items.

The figure also shows a smaller proportion of JOF > 0 for items in which the voice changed across study and test episodes. The difference between same and different-voice items shows that voice information is in fact encoded in memory and can affect recognition performance. However, the response curves for the different-voice items remained flat, rather than decreasing in slope. This indicates that as presentation frequency increased, subjects' ability to reject similar test items as new did not improve. The same pattern is found in the plural instruction condition (see Figure 4, lower panel).

239

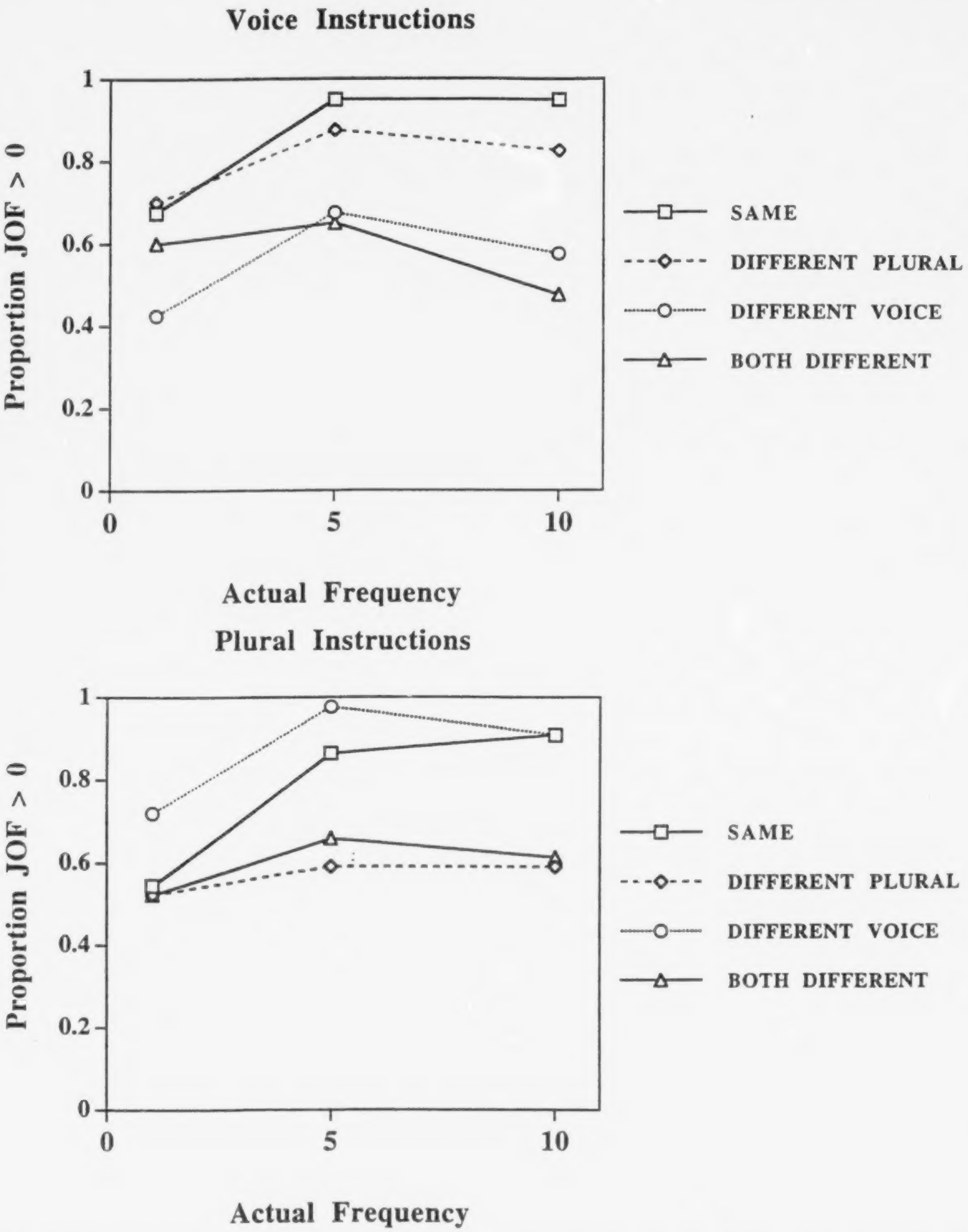


Figure 4. Proportions of judgments of frequency (JOFs) greater than zero are displayed as a function of presentation frequency and test item. The upper panel displays proportions for subjects in the voice instructions condition. The lower panel displays proportions for subjects in the plural instructions condition.

Table 1.

d' for Item Types and Test Condition in Experiment 1.

	Frequency at Study		
Item Type and Test Condition	One	Five	Ten
Same VOICE	1.20	2.12	2.12
Same PLURAL	.48	1.53	1.68
Different plural VOICE	1.30	1.88	1.71
Different plural PLURAL	.41	.63	.63
Different voice VOICE	.40	1.22	.89
Different voice PLURAL	1.08	1.91	1.68
Both different VOICE	.97	1.14	.56
Both different PLURAL	.41	.86	.71

Note: d' was derived from individual subject performance. Hits of 1.00 were truncated to .95; false alarms (FA's) of 0.00 were truncated to .05. FA's are based on "new" filler items; FA rate for the VOICE instruction condition = .28. FA rate for PLURAL instruction condition = .38.

The statistical analysis confirmed these patterns. Separate ANOVAs on the voice group and plural group data were conducted, with the within subject factors of frequency and item type. For subjects in the voice group, there was a main effect of presentation frequency [F(2, 38) = 4.61, p < .01]. However, post-hoc tests comparing performance across frequency conditions for each item type revealed that the effect of frequency arose entirely from differences between frequency = 1 and frequency = 5 [t(79) = 2.88, p < .0003]). This indicates that there was no additional learning of voice or plural information beyond frequency = 5. There was also a main effect of item type [F(3, 57) = 17.34, p < .0001]. In general, items that preserved voice across study and test were better recognized than items that changed voice. Same items were recognized better than different-voice [t(59) = 5.24, p < .0001.] and both-different items [t(59) = 5.29, p < .0001]. Different-plural items were recognized better than different-voice [t(59) = 5.01, p < .0001] and both-different items [t(59) = 4.09, p < .0001]. There was no interaction between frequency and item type in either group.

For subjects in the plural group, there was an effect of frequency [F(2, 80) = 9.42, p < .0004]. Post-hoc tests confirmed that there was no additional improvement in subjects learning of the same perceptual form-different perceptual form distinction beyond frequency = 5. The effect of item type was significant [F(3, 63) = 11.02, p < .0001]. Recognition was more accurate on same items than on different-plural items [t(65) = 3.40, p < .001] or both-different items [t(65) = 2.75, p < .0001]. Different-voice items were recognized more accurately than same items [t(65) = 2.34, p < .02], different-plural items [t(65) = 5.22, p < .0001], or both-different items [t(65) = 5.49, p < .0001]. There were no other effects.

In sum, the word recognition analysis revealed that when frequency judgments were recoded as hits, the benefits of a perceptual match across study and test were evident in both groups, and directly related to the test instructions: Subjects who were told to focus on voice information showed an effect of voice match; subjects who were told to focus on plurality information showed an effect of plurality match. In both groups, subjects largely ignored the irrelevant stimulus dimension when making their same form-different form discriminations. The fact that the encoding instructions were identical across both conditions suggests that the use of perceptual information during explicit retrieval is not automatic, but is moderated by the task demands and retrieval intentions of the subject.

As a whole, the results obtained in this experiment replicate and extend the earlier findings of Hintzman et al. (1992). The hallmarks of the "registration without learning" effect are an increased in the frequency judgments for targets and similar distractors, coupled with attenuated learning of the feature necessary to discriminate targets from similar distractors. In the present experiment, both patterns were present. Subjects knowledge that a word occurred continued to increase as presentation frequency increased, whereas their ability to reject similar distractors did not increase proportionally.

We also found no difference between voice learning and plurality. This finding was not expected. We predicted voice to be preserved to a greater extent than plurality because voice information is acoustically and perceptually richer and more complex, and because voices and words are processed in integral, parallel-contingent fashion. This latter fact suggested the possibility that the word information would "carry" voice information, and lead to proportional learning of both dimensions. The data showed instead that voice learning, like plural learning, stabilized after the first few repetitions. In Experiment 2, we further evaluate the retention of voice and plural information using several new procedures.

Although there is ample evidence that many acoustic-phonetic details are represented in memory along with more abstract phonological and semantic information (see Pisoni, 1993 for a review), this evidence is derived from experiments that examine the effects of stimulus variability over a single word repetition. It is important to point out that the presence of perceptual specificity effects in memory does not necessarily mean that perceptual information is represented completely and without error. To the contrary, the research findings described in Experiment 1 suggest that it is likely that such knowledge is incomplete and fragmentary. The results also show that the extent to which episodic information about the perceptual form of a word is encoded varies across instances of an item. Indexical information about a talker's voice, or acoustic-phonetic information about plurality, is most likely to be encoded and stored in memory when a word is novel, rather than highly familiar, and tends to largely remain constant across repetitions.

Experiment 2

Our first experiment revealed a dissociation between knowing that a word occurred and knowing details about the perceptual form of the word. A question to ask, then, is whether the dissociation occur because registration is truly without learning, or if instead, the results are merely a consequence of the retrieval task. In particular, the frequency judgment test required subjects to make two kinds of decisions simultaneously: Judge the familiarity of the item, and recollect perceptual information. In this way, the task incorporates both automatic and intentional retrieval processes. These processes can operate synergistically, or they can act in opposition, such as when a subject rejects a familiar distractor because they recall information about the target item (Jacoby, 1991). It is possible that greater learning of perceptual information would be apparent when voice or plurality discrimination is assessed separately from memory for frequency.

The first objective of Experiment 2 was to separate the effects of knowing a word occurred (regardless of its form) from the effects of knowing format-specific details about the word. Subjects were allowed to ignore study-to-test changes in the perceptual form of a test item when making their frequency judgments. We assessed memory for voice and plural information independently by asking subjects if the test word was presented in the same or different voice or plurality as the study item. The second objective of Experiment 2 was to compare explicit memory for voice and plurality information. Recall that a null effect of item type on frequency judgments was found among subjects in the plural instruction group (e.g., Figure 3). That measure also showed no effect of plural change on frequency judgments among subjects in the voice instruction group. Experiment 2 assessed whether these effects were the result of differences in memory for voice and plurality information.

To ensure that differences across experiments were due solely to retrieval processes and not encoding processes, the study instructions used in Experiment 2 were identical to those used in Experiment 1. Participants simply listened to the words on the study list with the expectation that they would be tested on the presentation frequencies of the words. After the study task, the participants answered four questions. The first two questions were confidence ratings and frequency estimates. Each provides a converging measure of subjects' ability to retrieve *word type* information. Subjects were told to give a positive confidence rating or frequency judgment even if the voice or plurality of the item had changed across study and test. The remaining two questions assessed subjects' ability to retrieve *token-specific* information. Subjects were asked to decide if the test word was in the same voice and the same plurality as the studied word. This allowed us to compare voice and plurality information by determining the extent to which each source is encoded and represented in memory. We hoped that the combination of these four questions would allow us to separate the effects that frequency and similarity have on the retrieval of type and token information.

Method

Participants

The participants were 20 Indiana University students who volunteered for the experiment in exchange for course credit. All were native speakers of English and reported no history of speech or hearing disorders. The participants were tested in groups of four or fewer. Four subjects were excluded from the analysis for failing to follow directions.

Apparatus

The stimulus materials were presented using an IBM-compatible personal computer. The words were presented binaurally to subjects at 75 dB SPL over matched and calibrated stereophonic headphones (BeyerDynamic DT-100).

Stimulus Materials

The stimulus materials were identical to those used in Experiment 1. In the study phase, participants heard a list of 205 successively presented words. The list contained 24 singular and plural target items produced by either a male or female talker. Target items were presented once or repeated five or ten times in the list. The study list also contained 77 nonrepeating fillers. In the test phase, participants heard a list of 48 successively presented words. Half were new items, and half were targets. There were four types of target items: "same," "different-voice," "different-plural," and "both-different."

Procedure

The study instructions were identical to those used in Experiment 1. Subjects were told to listen carefully to each item because they would be tested on the presentation frequencies of the words. Following the study phase, a test form was distributed. Participants gave four responses for each test word, taking as much time as necessary for each.

Questions 1 and 2 consisted of a confidence rating and a frequency judgment. Both questions assessed subjects' ability to retrieve *word type* information. Question 1 was a confidence rating on a -3 to +3 scale of the certainty that some form or version of the test word had been heard during the study phase. The following example was provided:

> *There are some words on the test list that were also on the study list in a different form. That is, the word may now be in a different plural form, or in a different voice, but was nevertheless present on the study list in some form. For example, you may hear "CAT" spoken in the female voice on the test, and remember that "CAT" was present in the study list, only it was spoken by a male voice. In this case, you should circle a number on the positive side of the response scale (+1, +2 or +3), depending on how certain or sure you are that you heard some form of CAT on the study list. If you don't remember the word, you should circle a number on the negative side of the response scale (-1, -2 or -3), depending on how certain or sure you are that you did not hear some form of the word on the study list.*

Question 2 was a frequency judgment in which subjects were asked to estimate the frequency of occurrence for each test word type. Subjects were instructed to circle a number from 1 to 15 if any version of the word was studied, and to circle zero if the word type was not studied.

Questions 3 and 4 were source judgments which assessed subjects' ability to retrieve *token-specific* information. Specifically, subjects were asked to decide if the test word was in the same voice and the same plurality as the studied word. Subjects made their voice judgments by checking either "male voice" or "female voice" on a response sheet. Similarly, they made their plurality judgments by checking either "singular" or "plural" on a response sheet.

Subjects were instructed to answer each question, regardless of their responses to the previous questions, and to guess if they were unsure. The order of questions 3 and 4 were counterbalanced, as was the order of the response alternatives within each question. Subjects were carefully instructed as to how to answer each of the four test questions, and were given several examples to ensure that they understood the test tasks. The entire experimental session lasted less than 30 minutes, after which the subjects were debriefed.

Results

Frequency Judgments

Mean Frequency Judgments. Because the histograms revealed a unimodal frequency distribution for each item type, the analysis of the mean frequency estimates included all four item conditions. Subjects showed sensitivity to differences in the study presentation frequencies [F(2, 38) = 39.45, p < .0001]. Mean JOFs across the three frequency levels were 1.5, 5.7, and 9.1, which reveals a close relationship between judged frequency and actual frequency. However, there was no effect of item type, and it did not interact with frequency.

Nonzero JOFs. As in Experiment 1, an analysis was conducted on the nonzero JOFs in order to determine if frequency judgments made to similar items were lower than judgments made to same items. This analysis excluded four subjects who failed to contribute at least one nonzero judgment for each item type at each frequency. The analysis of the nonzero JOFs (see Figure 5) revealed a highly significant effect of frequency [F(2, 22) = 70.53, p < .0001]. There were no other effects.

Insert Figure 5 about here

Word Recognition. Frequency judgments were collapsed into a binary recognition measure. Recognition of a word type as "old" was defined as a positive frequency judgment. Figure 6 shows the proportion of correct recognitions as a function of presentation frequency and item type (see Table 2 for the corresponding *d'* values). The figure shows that word type recognition improved with presentation frequency [F(2, 30) = 40.20, p < .0001]. In fact, performance is close to ceiling after the fifth presentation. Such high performance levels were unanticipated given that the methodology of this experiment is very similar to Experiment 1, where ceiling effects were not found.

Another difference between the present experiment and Experiment 1 is that there was no effect of item type in the present experiment. This indicates that recognition memory for word types was not affected by the physical similarity between the study tokens and test tokens in this task, and confirms that subjects were following directions.

Insert Figure 6 about here

Confidence Judgments

Mean Frequency Judgments. Mean confidence ratings are displayed in Figure 7 as a function of presentation frequency and item type. An analysis was conducted on the confidence judgments in order to determine if confidence increased with study repetitions and if different form items were given lower confidence ratings than same form items. There was, of course, a highly significant effect of frequency [F(2, 22) = 70.53, p < .0001], indicating that subjects' confidence that a word type was "old" increased dramatically with presentation frequency. The effect of item type was not significant, which indicates that recognition of same and similar items was accomplished with equal confidence. Frequency and item type did not interact.

Insert Figure 7 about here

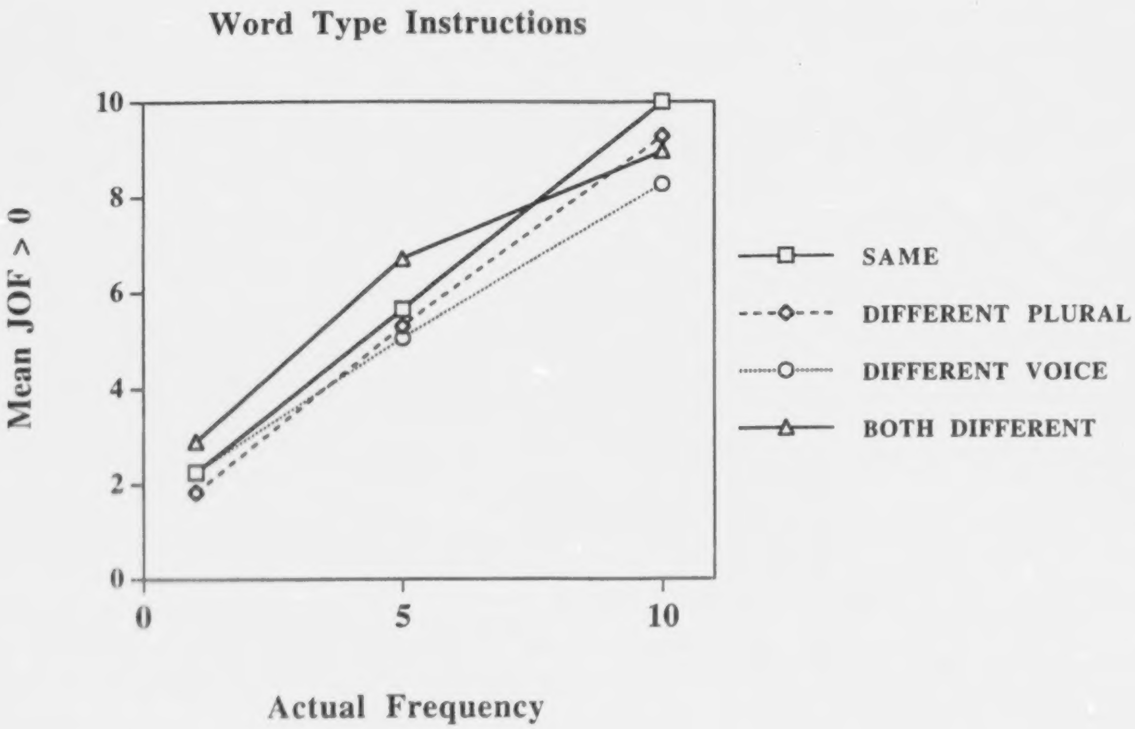


Figure 5. Mean judgments of frequency (JOFs) greater than zero are displayed as a function of presentation frequency and test item.

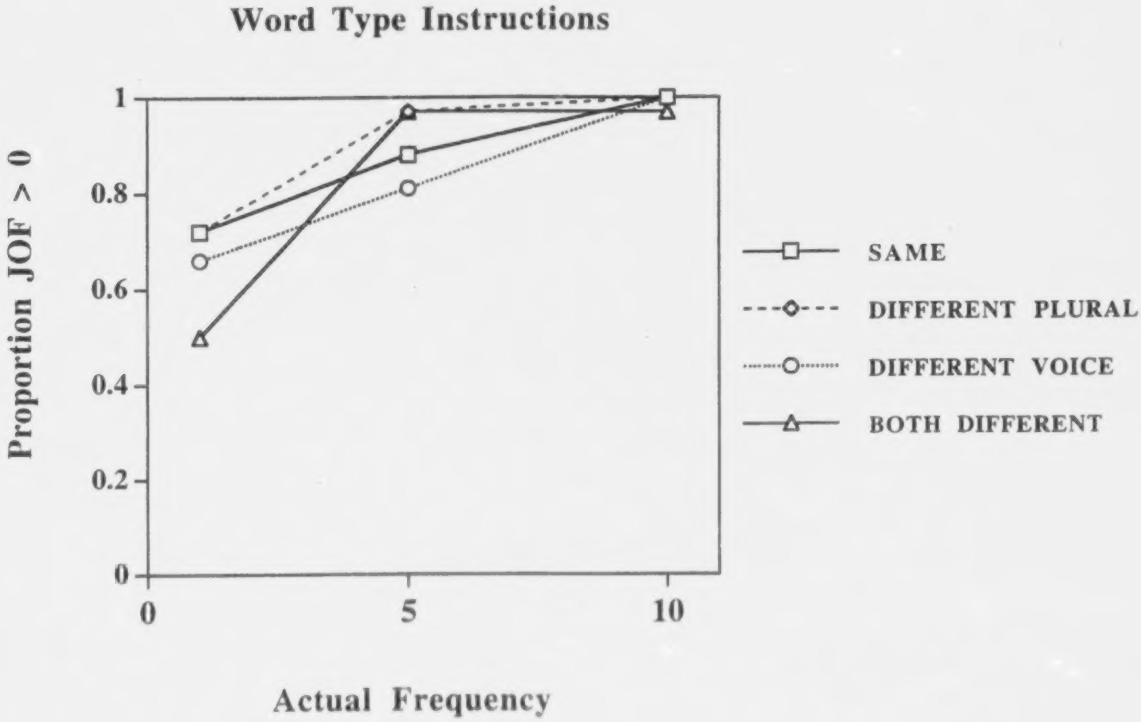


Figure 6. Proportions of judgments of frequency (JOFs) greater than zero are displayed as a function of presentation frequency and test item type.

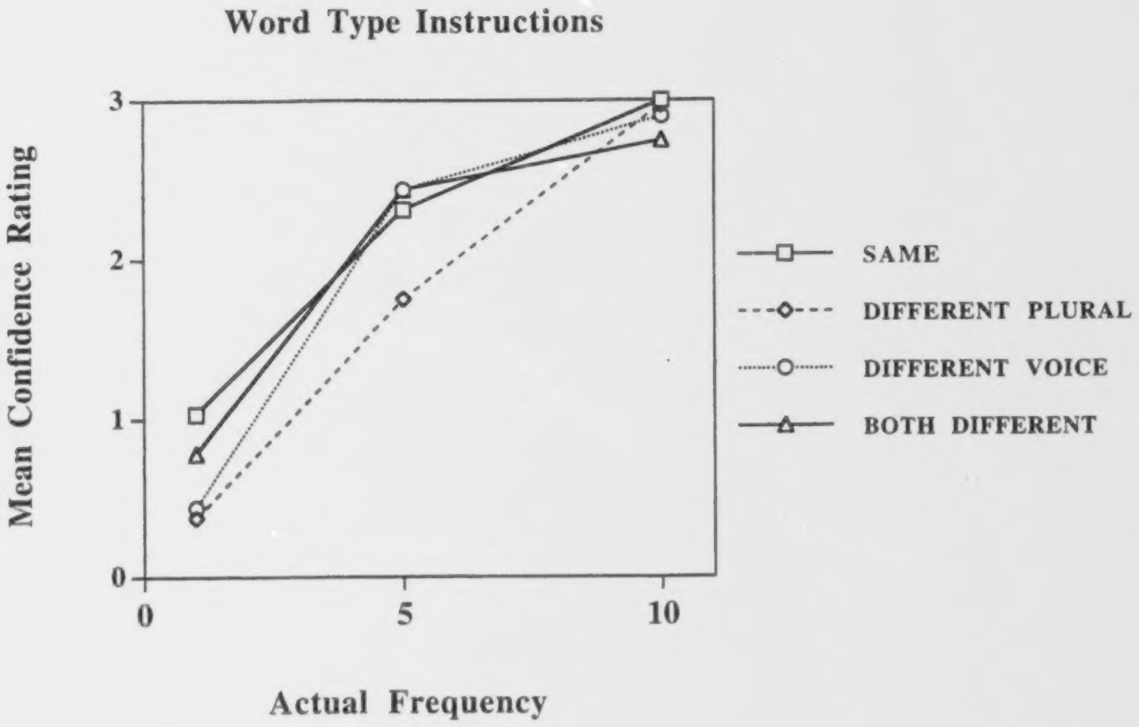


Figure 7. Mean confidence ratings are displayed as a function of presentation frequency and test item type.

Table 2.

d' for Item Types and Test Condition in Experiment 2.

	Frequency at Study		
Item Type and Test Condition	One	Five	Ten
Same JOF	1.11	1.52	1.93
Same CONF	1.26	1.67	2.08
Different plural JOF	.90	1.21	1.93
Different plural CONF	1.05	1.34	2.08
Different voice JOF	1.10	1.10	1.93
Different voice CONF	.85	1.98	2.08
Both different JOF	.90	1.82	1.72
Both different CONF	1.05	1.98	1.88

Note: "CONF" = recognition data from the confidence judgments. "JOF" = recognition data from the judgments of frequency. d' was derived from individual subject performance. Hits of 1.00 were truncated to .95; false alarms (FA's) of 0.00 were truncated to .05. FA's are based on "new" filler items. FA rate for JOF = .40. FA rate for CONF = .34.

Word Recognition. Confidence judgments were also reduced to an old/new recognition measure. Positive confidence ratings (1, 2, or 3) were recoded as "old" responses, whereas negative confidence ratings (-1, -2, or -3) were recoded as "new" responses. The proportion of positive confidence responses increased systematically with presentation frequency [$F(2, 30) = 45.33$, $p < .0001$] but did not differ among the test items. There were no other effects.

So far, the data from several measures convey the same story: The more often a given word was encountered at study, the higher the frequency estimate was for that item at test. Similarly, frequently encountered words were also more likely to be recognized, and the recognition judgments were accompanied by greater certainty that the item was old. Accordingly, these findings reflect the expected effect of study repetition on explicit memory performance.

The recognition data derived from the confidence judgments were nearly identical to the recognition data derived from the frequency judgments. Of course, differences in the two measures are partly obscured by the very high performance levels at frequency = 10. Nevertheless, subjects appear to have used the two response scales similarly, despite the fact that the frequency scale has twice the number of response categories (Proctor, 1977). Although the high performance levels in the present experiment do not allow us to definitively conclude that the same processing mechanism underlies judgments of presentation frequency and recognition memory, our results are consistent with many other experiments that do subscribe to this view (e.g., Harris et al., 1980; Hintzman, 1988).

In addition, word type recognition accuracy and confidence were the same across the four item types. Presumably, token-specific information is irrelevant to a task that requires the retrieval of word type information. That is, performance was unaffected by study-to-test changes in voice and plurality because subjects were asked to ignore such changes and responded accordingly. This led to a large number of positive JOF's accompanied by high levels of confidence that the item types were present on the study list.

Source Judgments

To analyze the explicit voice and plural source judgments, the proportion of correct same and different responses were calculated for trials in which the word was correctly recognized. Figure 8 displays the proportion of correct voice and plural judgments as a function of frequency. The figure shows that explicit memory for voice and plural information was overall above chance and slightly increased with study frequency.

Two statistical analyses were performed on the source judgments. One analysis conditionalized the judgments on correct recognition from the confidence ratings, whereas a second analysis conditionalized the judgments on correct recognition from the frequency judgment task. Since both analyses led to the same conclusions, only the statistical analysis based on the confidence ratings is presented here.

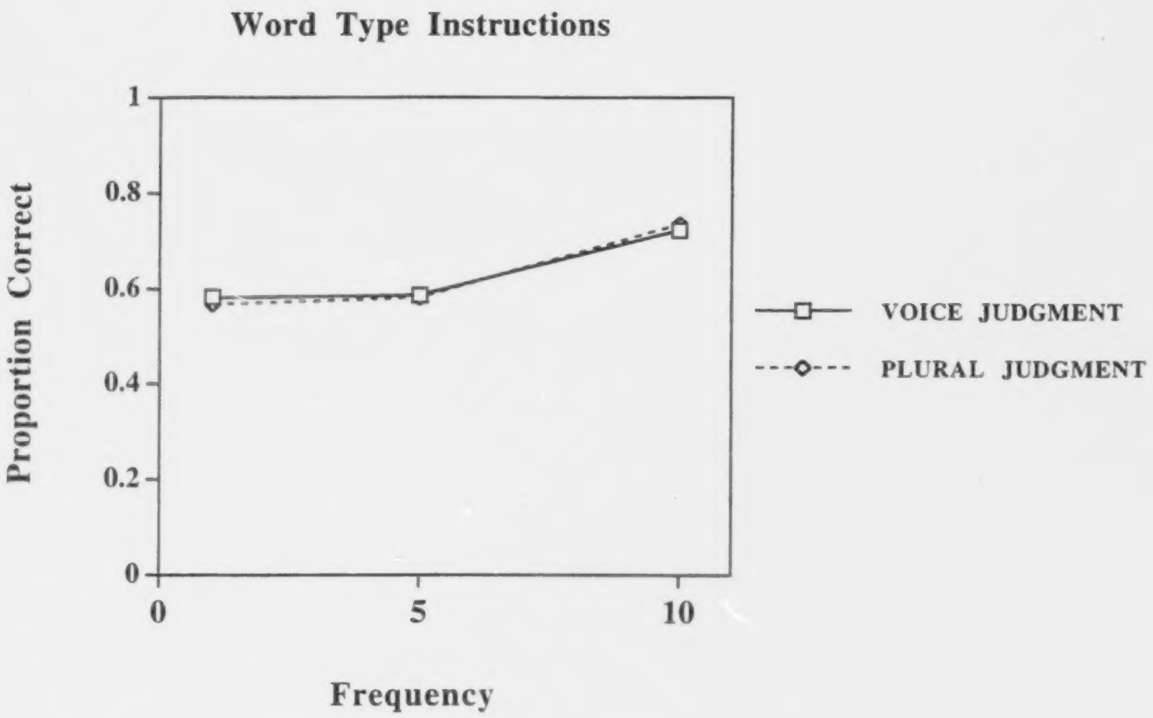
Insert Figure 8 about here

An ANOVA with the factors source judgment (voice or plural) and presentation frequency (1, 5, 10) showed a significant effect only for frequency [$F(2, 60) = 12.67$, $p < .0001$]. Accuracy was the same for the plural and voice source judgments (overall proportion correct was .63 for both judgments). Planned t-tests comparing voice recognition at each frequency level showed that overall voice recognition performance was significantly different from chance in all three frequency conditions ([$t(15) = 2.68$, $p < .01$] for frequency = 1; [$t(15) = 3.87$, $p < .001$] for frequency = 5; [$t(15) = 6.54$, $p < .0001$] for frequency = 10). In contrast, performance in discriminating same and different-plural items across study and test exceeded chance only at frequency =10, [$t(15) = 9.46$, $p < .0001$].

The source judgments directly assessed the encoding and retrieval of token-specific perceptual information. The memorability of voice and plurality information was equal, and source judgment accuracy showed only a small improvement even after 10 repetitions of a word-voice/plurality pair. The fact that there was an increase (albeit a small one) in source judgment accuracy as a function of frequency shows that registration is not *entirely* without learning. However, the amount of learning is modest, and certainly not proportional to word learning. Taken together, the explicit source judgments from this experiment and the discrimination data from Experiment 1 lead to the same conclusion: The enhanced ability to recognize a word that has been presented many times does not guarantee that the perceptual information is represented with any more detail than a word presented only once.

General Discussion

The present research sought to extend the "registration without learning effect" (Hintzman et al., 1992) to a new domain, for a new and more complex feature and with several new procedures. To accomplish this, we had subjects listen to a list of words spoken by a male or female talker, and presented



Word Type Instructions

Figure 8. Proportion of correct voice and plural explicit source judgments are displayed as a function of presentation frequency.

in singular or plural form. Target items were presented various times in the study list. In Experiment 1, subjects estimated the frequency with which a word occurred at study, excluding items that were in a different voice from the study item (voice instructions group) or in a different pluralization from the study item (plural instructions group). In Experiment 2, the test task required subjects to estimate the frequency with which a word type occurred at study, ignoring differences in voice or plurality. Subjects also provided confidence ratings and made explicit voice and plurality judgments about the test words. In both experiments, the test list contained items that differed in their physical similarity to the study items by one or two dimensions (voice, plurality or both). Both experiments measured the effects of familiarity and similarity on subjects' memory for frequency and their ability to make fine-grained distinctions among similar items.

The major findings were as follows: (1) Frequency judgments, recognition accuracy, and confidence ratings for word tokens and word types were higher for words repeated many times during study as compared to words presented just once. (2) In contrast, repeating a word many times during study did not improve subjects' memory for the perceptual details associated with the word. (3) The asymptotic learning function for the feature that discriminates same form items from similar distractors was not specific to the frequency judgment task used in Experiment 1. Explicit source judgments from Experiment 2 also revealed attenuated learning. (4) The complexity of the stimulus feature did not affect learning. (5) The effect of the similarity between study and test items was dependent on instructions at retrieval. Together, these findings demonstrate that the acquisition of instance-specific perceptual knowledge varies across word repetitions.

Repetition and Perceptual Processing

The present results are consistent with the idea that the gains in word recognizability that occur when an item is repeated are attributable to the listener's failure to engage in bottom-up processing of surface information. The idea here is that perceptual processing is suppressed for inputs that match expectations and accentuated for inputs that are unexpected or novel (Hintzman et al., 1992; Johnston & Hawley, 1994). For example, in the present experiments, when listeners encountered a word that was presented earlier in the study list, they presumably realized (either consciously or unconsciously) that they had heard the word earlier in the list, and consequently, didn't allocate processing resources reaffirming details about a word they already knew. The perceived familiarity of the item served to disguised the fact that their knowledge about the structure of the word was incomplete. Consequently, the subject later had difficulty distinguishing between the studied word and a similar distractor because attention to perceptual features had largely been truncated after the first few repetitions (DiGirolamo & Hintzman, 1997; Tulving & Kroll, 1997).

In the present study, variation in voice or plurality occurred *across* study and test phases of the experiment. It is worth noting that DiGirolamo and Hintzman (1997) have recently obtained similar effects when the attributes of a repeated object are varied *within* the study phase. In their study, DiGirolamo and Hintzman varied the perceptual form of an object either early or late in the study list. For some objects, the orientation of the first presentation differed from presentations 2-5, whereas for other objects, the fifth presentation differed from presentations 1-4. Subjects were then given a forced choice discrimination test which assessed their ability to remember whether an item occurred in one orientation or both orientations. DiGirolamo and Hintzman found that subjects reported seeing both orientations more often if orientation changed early in the list rather than later in the list. They argue that their findings are not the result of a simple primacy effect, but stem from the fact that information processing on early repetitions of an item was qualitatively different from information processing on later repetitions of the same item.

An alternative perspective from which to view the effects of repetition is to suppose that repetition serves to build a "unitized" response code (Feustel, Shiffrin & Salasoo, 1983; Shiffrin & Lightfoot, 1997). Unitization refers to a process that integrates co-occurring parts of an event into a single functional unit. Unitization can be conceived of as an abstraction process whereby features that are not relevant to the task are less likely to become integrated. Unitization allows a complex stimulus to be identified based on the most relevant or diagnostic features. However, once an item is unitized, it becomes more difficult to detect incidental or features on subsequent repetitions. Thus, unitization may play a role in accounting for the registration without learning phenomena.

The unitization view also suggests that perceptual information, such as information about a talker's voice, may continue to be encoded across repetitions if it is perceived as being diagnostic of a word's identity. For example, imagine if each of the 24 target words used in the present experiment were produced by one of 24 different voices. If a specific voice reliably co-occurred with just one word within a list, listeners may detect the relevance of the voice as a retrieval cue. Consequently, voice information may become integrated with the word in such a way that promotes learning of voice information across repetitions. We are currently exploring this possibility.

Of course, these accounts of registration without learning rests entirely on data derived from explicit memory tasks. Explicit tasks are largely driven by the match between conceptual or contextual information at study and test and tend to be unaffected by perceptual information. Implicit tasks tap different kinds of cognitive operations than do explicit tasks and thus, recover different information (for a review, see Roediger & McDermott, 1993; Schacter, 1987). In particular, implicit test performance tends to be facilitated by perceptual consistency across study and test. It is possible that perceptual information associated with a familiar item is represented veridically, but that the experimental task used in the present experiments provided only a limited window on the underlying information. Further research is needed in order to determine if the effects reported here and those reported by Hintzman and colleagues are less about "registration without learning" and more about "learning without retrieval".

Retrieval Models

The results of the present experiments underscore the importance of a retrieval model based on two processes (familiarity and recall). The need for two processes is necessary because discriminative responding of the kind required by the frequency task used in Experiment 1 could not be mediated by a unidimensional familiarity signal. A second mechanism is needed that supports the recollection of instance-specific details from a prior episode, and allows one to know that a highly familiar item was not actually on the study list.

Further empirical support for the distinction between familiarity and recall comes from a recent experiment that used a response-signal paradigm (Hintzman & Caulton; 1997; cf. Hintzman & Curran, 1994; Mulligan & Hirshman, 1995). This method is an elegant way to explore how the retrieval of a word unfolds over time. Hintzman and Caulton found that the retrieval dynamics of familiarity and recall can be differentiated by their speed-accuracy retrieval patterns and by their sensitivity to the effects of repetition and similarity. Although a few qualitative models (Johnston, Jacoby & Dark, 1975; Mandler, 1980) and formal quantitative models of memory include familiarity and recall processes (e.g., Minerva 2, Hintzman, 1988; REM, Shiffrin & Steyvers, 1997; SAM, Gillund & Shiffrin, 1984), the extent to which recall contributes to recognition memory and frequency judgments is currently not well specified. This timely and interesting topic is worthy of further research and simulation.

Encoding of Talker Information

The findings from Experiments 1 and 2 have implications for the manner in which we characterize the processing or "normalization" of voices. Recent research demonstrates that familiarizing subjects with the voices in which speech stimuli are presented improves word recall (Lightfoot, 1989) and speech intelligibility (Nygaard, Sommers & Pisoni, 1994). For example, listeners who learn to identify a set of talkers from sentence length utterances are able to transcribe new sentences presented in white noise more accurately than subjects who are unfamiliar with the talkers (Nygaard & Pisoni, 1995). These experiments clearly show that familiarity with a voice has a relationship to the ongoing analysis of speech. Nygaard & Pisoni (1995) argue that the increase in sensitivity to talker-specific information brought about by training automatically increases sensitivity to the linguistic information in the signal.

Their results are grounded in the proceduralist model described by Kolers (1979; Kolers & Roediger, 1984), in which the pattern-analyzing operations or "procedures" involved in encoding an item serve as the basis for the memory representation of the event. These procedures are invoked to supplement the analysis of a repeated event. The fact that pattern-analyzing operations do not have to be constructed anew means that familiar items, or new items presented in a familiar script or voice, can be identified more efficiently. An important, albeit implicit, assumption that underlies this view is that the effect of practice is to change *how* information is processed, rather than *which* information is processed. Listeners attend to, encode, and process the same amount of information regardless of their familiarity with the stimulus item. Accordingly, the gain in the efficiency of processing does not imply that the processing of perceptual information is in any way curtailed or attenuated.

However, the experiments in this report, as well as those by Hintzman and his colleagues raise the possibility that the improvement in spoken language processing resulted not from *better* use of voice information, but from *less* use of voice information. That is, the effect of training and practice may have been to induce a systematic reduction in the amount of perceptual information processed, rather than to change the efficiency with which perceptual analysis was performed.

The notion that perceptual details are less memorable if they occur on later repetitions may also shed light on the interaction between talker variability and list position (Martin, Mullennix, Pisoni & Summers, 1989), on differences in the encoding and retrieval of single-talker word lists as compared to multiple-talker lists (Mullennix, Pisoni & Martin, 1989), and on the mechanisms assumed to underlie speaker normalization (see Pisoni, 1996). We claim that what is commonly referred to as "speaker normalization" may simply be a general epiphenomenon of redistributed perceptual processing or learned inattention that commonly occurs in response to many objects and events, rather than a specific linguistic process dedicated to the recovery of canonical phonetic units from speech.

Repeated exposure may also affect the influence of speaker information on other tasks. For instance, Walker, Bruce & O'Malley, (1995) found that subjects who are familiar with the face producing an audiovisual syllable are less susceptible to influences of facial speech information on auditory speech perception. That is, familiar faces produce weaker McGurk effects. The relative ease with which faces become familiar as compared to voices may also play a role in the weak effects of face variation relative to voice variation on spoken word recognition (Sheffert & Fowler, 1995). Further research is needed to understand whether familiarity directly affects the use of talker-specific information during bimodal speech processing.

Conclusion

The experiments reported here examined the perception and encoding of instance-specific perceptual information as a function of the amount of prior exposure a listener has with a particular item. The results replicate and extend the prior findings of Hintzman, et al. (1992) by revealing a dissociation between knowing that a spoken word occurred and knowing perceptual details about the word. The collective import of the studies is to show that "registration without learning" occurs for auditorily presented words, for simple and complex stimulus dimensions, and across various experimental settings. The findings add to a growing literature on the factors that are important in determining what knowledge a subject encodes during spoken word recognition, and help to delimit theoretical interpretations of memory.

References

Bradlow, A. R., Nygaard, L. C. & Pisoni, D. B. (in press). Effects of talker, rate and amplitude variation on recognition memory for spoken words. *Perception and Psychophysics*.

Bricker, P. D., & Pruzansky, S. (1976). Speaker recognition. In N.J. Lass (Ed.), *Contemporary Issues in Experimental Phonetics* (pp. 295-326). New York: Academic Press.

Craik, F. I. M. & Kirsner, K. (1974). The effects of speaker's voice on word recognition. *Quarterly Journal of Experimental Psychology*, **26,** 274-284.

DiGirolamo, G. & Hintzman, D. L. (1997). First impressions are lasting impressions: A primacy effect in memory for repetitions. *Psychonomic Bulletin and Review*, **4**, 121-124.

Feustel, T. C., Shiffrin, R. M. & Salasoo, A. (1983). Episodic and lexical contributions to the repetition effect in word identification. *Journal of Experimental Psychology: General*, **112**, 309-346.

Flexser, A. J. & Bower, G. H. (1975). Further evidence regarding instruction effects on frequency judgments. *Bulletin of the Psychonomic Society*, **6**, 321-324.

Geiselman, R. E. & Bellezza, F. S. (1977). Incidental retention of speaker's voice. *Memory and Cognition*, **5**, 658-665.

Gillund, G. & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review*, **91**, 1-67.

Green, K. P., Tomiak, G. R. & Kuhl, P. K. (1997). The encoding of rate and talker information during phonetic perception. *Perception and Psychophysics*, **59**, 675-692.

Greene, R. L. (1984). Incidental learning of event frequency. *Memory and Cognition*, **12**, 90-95.

Harris. G., Begg, I. & Mitterer, J. (1980). On the relation between frequency estimates and recognition memory. *Memory & Cognition*, **8**, 99-104.

Hasher, L. & Zacks, R. T. (1974). Automatic and effortful processes in memory. *Journal of Experimental Psychology: General*, **108**, 356-388.

Hasher, L. & Zacks, R. T. (1984). Automatic processing of fundamental information. *American Psychologist*, **12**, 1372-1388.

Hintzman, D. L. (1969). Apparent frequency as a function of frequency and the spacing of repetitions. *Journal of Experimental Psychology*, **80**, 139-145.

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace model. *Psychological Review*, **95**, 528-551.

Hintzman, D. L. & Block, R. A. (1971). Repetition and memory. Evidence for a multiple-trace hypothesis. *Journal of Experimental Psychology*, **88**, 297-306.

Hintzman, D., Block, R. A. & Inskeep, N. R. (1972). Memory for the mode of input. *Journal of Verbal Learning and Verbal Behavior*, **11**, 741-749.

Hintzman, D. L. & Caulton, D. A. (1997). Recognition memory and modality judgments: A comparison of retrieval dynamics. *Journal of Memory and Language*, **37**, 1-23.

Hintzman, D. L. & Curran, T. (1994). Retrieval dynamics of recognition and frequency judgments: Evidence for separate processes of familiarity and recall. *Journal of Memory and Language*, **33**, 1-18.

Hintzman, D. L. & Curran, T. (1995). When encoding fails: Instructions, feedback, and registration without learning. *Memory and Cognition*, **23**, 213-226.

Hintzman, D. L., Curran, T. & Oppy, B. (1992). Effects of similarity and repetition on memory: Registration without learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **18**, 667-680.

Howell, W. C. (1973). Storage of events and event frequencies: A comparison of two paradigms in memory. *Journal of Experimental Psychology*, **98**, 260-263.

Humphreys, M. S., Bain, J. D. & Pike, R. (1989). Different ways to cue a coherent memory system: A theory for episodic, semantic, and procedural tasks. *Psychological Review*, **96**, 208-233.

Jacoby, L. L. (1991). A Process Dissociation Framework: Separating automatic from intentional uses of memory. Journal of Memory and Language, **30**, 513-541.

Johnston, W. A., Dark, V. J., & Jacoby, L. L. (1985). Perceptual fluency and recognition judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **11**, 3-11.

Johnston, W. A. & Hawley, K. J. (1994). Perceptual inhibition of expected inputs: The key that opens closed minds. *Psychonomic Bulletin and Review*, **1**, 56-72.

Kolers, P. A. (1979). A pattern-analyzing basis for recognition memory. In L. S. Cermak & F. I. M. Craik (Eds.) *Levels of processing and human memory*. Hillsdale, NJ: Erlbaum.

Kolers, P. A. and Roediger (1984). Procedures of the mind. *Journal of Verbal Learning and Verbal Behavior*, **23**, 425-449.

Kraut, A. G. (1976). Effects of familiarization on alertness and encoding in children. *Developmental Psychology*, **12**, 491-496.

Kraut, A. G., Smothergill, D. W. & Farkas, M. S. (1981). Stimulus repetition effects on attention to words and colors. *Journal of Experimental Psychology: Human Learning and Memory*, **7**, 1303-1311.

Lightfoot, N. (1989). Effects of talker familiarity on serial recall of spoken word lists. In *Research on Speech Perception Progress Report No.* **15**. (pp. 419-444). Bloomington, IN: Speech Research Laboratory, Indiana University.

Mandler, G. (1980). Recognizing: The judgment of previous occurrence. *Psychological Review*, **87**, 252-271.

Martin, C. S., Mullennix, J.W., Pisoni, D.B. & Summers, W. V. (1989). Effects of talker variability on recall of spoken word lists. *Journal of Experimental Psychology: Human Learning and Memory*, **15**, 676-684.

Metcalfe, J. (1982). A composite holographic associative recall model. *Psychological Review*, **89**, 627-661.

Mullennix, J.W., & Pisoni, D.B. (1990). Stimulus variability and processing dependencies in speech perception. *Perception and Psychophysics*, **47**, 379-390.

Mullennix, J.W., Pisoni, D.B. & Martin, C. S. (1989). Some effects of talker variability on spoken word recognition. *Journal of the Acoustical Society of America*, **85**, 365-378.

Mulligan, N. & Hirshman, E. (1995). Speed-accuracy trade-offs and the dual process model of recognition memory. *Journal of Memory and Language*, **34**, 1-18.

Murdock, B. B. (1982). A theory for the storage and retrieval of item and associative information. *Psychological Review*, **89**, 609-626.

Naveh-Benjamin, M. & Jonides, J. (1986). On the automaticity of frequency coding: Effects of competing task load, encoding strategy, and intention. *Journal of Experimental Psychology: Learning, Memory and Cognition*, **3**, 378-386.

Nickerson, R. S. & Adams, M. J. (1979). Long-term memory for a common object. *Cognitive Psychology*, **11**, 287-307.

Nygaard, L.C., Sommers, M.S., & Pisoni, D.B. (1994). Speech perception as a talker-contingent process. *Psychological Science*, **5**, 42-46.

Nygaard, L.C. & Pisoni, D.B. (1995). Talker and task-specific perceptual learning in speech perception. In *Proceedings of the XIIIth International Congress of Phonetic Sciences* (pp. 194-197). Stockholm: Stockholm University.

Palmeri, T. J., Goldinger, S. D., & Pisoni, D. B. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology: Learning, Memory and Cognition,* **19**, 309-328.

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America,* **24**, 175-184.

Pisoni, D. B. (1993). Long-term memory in speech perception: Some new findings on talker variability, speaking rate and perceptual learning. *Speech Communication,* **13**, 109-125.

Pisoni, D. B. (1996). Some thoughts on "normalization" in speech perception. In K. Johnson & J. W. Mullennix (Eds.), *Talker variability in speech processing* (pp. 9-32). San Diego: Academic Press.

Posner, M. I. & Boies, S. (1971). Components of attention. *Psychological Review,* **78**, 391-408.

Proctor, R. W. (1977). The relationship of frequency judgments to recognition: Facilitation of recognition and comparison to recognition-confidence judgments. *Journal of Experimental Psychology: Human Learning and Memory,* **3**, 679-689.

Raaijmaker, J. G. & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review,* **88**, 93-134.

Richardson-Klavehn, A. & Bjork, R. A. (1988). Measures of memory. *Annual Review of Psychology,* **39**, 475-543.

Roediger, H. L. & McDermott, K. B. (1993). Implicit memory in normal human subjects. In H. Spinnler & F. Boller (Eds.), *Handbook of neuropsychology, Vol. 8.* (pp. 63-131). Amsterdam: Elsevier.

Rose, R. J. & Rowe, E. J. (1976). Effects of orienting task and spacing of repetitions of frequency judgments. *Journal of Experimental Psychology: Human Learning and Memory,* **2**, 142-152.

Rowe, E. J. (1974). Depth of processing in a frequency judgment task. *Journal of Verbal Learning and Verbal Behavior,* **13**, 636-643.

Schacter, D. L. (1987). Implicit memory: History and current status. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* **13**, 501-518.

Sheffert, S. M. (in press). Contributions of surface and conceptual information on spoken word and voice recognition. *Perception and Psychophysics.*

Sheffert, S. M. & Fowler, C. A. (1995). The effects of voice and visible speaker change on memory for spoken words. *Journal of Memory and Language,* **34,** 665-685.

Shiffrin, R. M. & Lightfoot, N. (1997). Perceptual learning of alphanumeric-like characters. In R. Goldstone, D. Medin & P. Schyns (Eds.), *Perceptual Learning* (pp. 45-84). New York: Academic Press.

Shiffrin, R. M. & Steyvers, M. (1997). A model of recognition memory: REM - Retrieving effectively from memory. *Psychonomic Bulletin & Review*, **4**, 145-166.

Torretta, G. M. (1995). The easy-hard word multi-talker speech database: An initial report. *Research on Spoken Language Processing, Progress Report No. 20*, (pp. 321-334), Indiana University, Bloomington, IN.

Treisman, A. (1992). Perceiving and re-perceiving objects. *American Psychologist*, **47**, 862-875.

Tulving, E. & Kroll, N. (1995). Novelty assessment in the brain and long-term memory encoding. *Psychonomic Bulletin & Review*, **2**, 387-390.

Walker, S., Bruce, V. & O'Malley, C. (1995). Facial identity and facial speech processing: Familiar faces and voices in the McGurk effect. *Perception and Psychophysics*, **57**, 1124-1133.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 21 (1996-1997)
Indiana University

Improvements in Speech Perception in Prelingually-Deafened Children: Effects of Device, Communication Mode, and Chronological Age[1]

Ted A. Meyer,[2] Mario A. Svirsky,[2] Karen I. Kirk,[2] & Richard T. Miyamoto[2]

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405

---

[1] This research was supported by NIH Research Grant DC-00012 and DC00064 to Indiana University. Portions of the paper were presented at the 20th Annual Meeting of the Association for Research in Otolaryngology, St. Petersburg Beach, FL, February, 1997. We would like to thank Amy M. Robbins, Susan T. Sehgal, and Allyson I. Riley for help in data collection, Theresa S. Kerr for help in data organization, and Linette A. Caldwell for clerical assistance. We would also like to thank David B. Pisoni for his comments on an earlier version of the manuscript.

[2] Indiana University School of Medicine, Department of Otolaryngology, DeVault Otologic Research Laboratory, Indianapolis, IN.

T.A. MEYER, M.A. SVIRSKY, K.I. KIRK AND R.T. MIYAMOTO

Improvements in Speech Perception in Prelingually-Deafened Children: Effects of Device, Communication Mode, and Chronological Age

Abstract. Miyamoto, Osberger, Todd, Robbins, Karasek et al. (1994) compared the speech perception skills of children with profound prelingual deafness who had received the Nucleus multichannel cochlear implant (CI) to those who were not implanted and used conventional hearing aids (HA). The CI users were tested over time and the HA users were tested at a single point in time. They found that the CI users improved their scores on speech perception tasks a great deal over time. After about 2.5 years of device use, the CI users were performing better than the average performance from a group of Silver (PTA=104 dB HL) HA users on all tests, and their scores were approaching the average scores from a group of Gold (PTA=94 dB HL) HA users except on tests of open-set sentence recognition.

The present investigation expanded on the earlier study of Miyamoto et al. by examining speech perception scores over time for both groups of children as a function of communication mode of the child. Separate linear regressions of speech perception scores as a function of age were computed to estimate the rate of improvement in speech perception abilities that might be expected due to maturation for the HA users (n=58) within each communication mode. The resulting lines were used to compare the estimated rate of speech perception growth for each HA group to the observed gains in speech perception made by the children with multichannel CIs. A large number of children using CIs (n=74) were tested over a long period of implant use (mean, 3.5 years; max, 8.5 years). In general, speech perception scores for the children using CIs were higher than those predicted for the Silver HA users, and they approached the scores predicted for the Gold HA users.

Introduction

Numerous research groups have examined speech perception performance in prelingually-deafened children with multichannel cochlear implants (Carney et al., 1991; Cowan et al., 1994; Fryauf-Bertschy, Tyler, Kelsay, & Gantz, 1992; Fryauf-Bertschy, Tyler, Kelsay, Gantz, & Woodworth, 1997; Gantz, Tyler, Tye-Murray, & Fryauf-Bertschy, 1994; Gantz, Tyler, Woodworth, Tye-Murray, & Fryauf-Bertschy 1994; Geers & Brenner, 1994; Kirk, Osberger, & Pisoni, 1995; Miyamoto, Osberger, Robbins, Myres, Kessler et al., 1991; Miyamoto, Osberger, Robbins, Myres, & Kessler 1993; Miyamoto, Osberger, Todd, Robbins, Karasek et al., 1994; Miyamoto, Osberger, Todd, Robbins, Stroer et al., 1994; Miyamoto, Kirk, Todd, Robbins, & Osberger, 1995; Miyamoto, Kirk, Robbins, Todd, & Riley, 1996; Osberger, Miyamoto et al., 1991; Osberger, Robbins et al., 1991; Somers, 1991; Staller, Beiter, Brimacombe, Mecklenburg, & Arndt, 1991; Staller, Dowell, Beiter, & Brimacombe, 1991; Waltzman, Cohen, & Shapiro, 1992; Waltzman et al., 1994; Waltzman et al., 1995). Early research in this field utilized descriptive analyses of the speech recognition scores of cochlear implant (CI) users. This method of reporting data provided an understanding of some of the benefits in speech perception the CI users received from their implants, but it did not provide any insight as to any gains these children might have made had they not been implanted. Some of the more recent studies have used a treatment/control group paradigm, where speech recognition scores of CI users have been compared to scores from unimplanted children who use hearing aids (HAs) or tactile devices to enhance their communication. Such studies have shown that the average scores of children with CIs exceed the average score of HA users who have unaided thresholds poorer than 100 dB HL (Osberger, Miyamoto et al., 1991; Somers, 1991). In general, one of two methods was used to make comparisons between groups: (a) the longitudinal performance of pediatric CI users measured over time was compared to the performance of a cross-section of children using HAs assessed at a single point in time (Miyamoto, Osberger, Todd, Robbins, Karasek et al., 1994); or (b) the scores from small groups of CI users were compared to scores from age-matched HA or tactile aid users who served as controls (Geers & Brenner, 1994; Kirk et al., 1995).

These methods are an improvement over simply describing and reporting the changes seen over time with device use, but generalizations from these types of studies are difficult to make. In the first method described (longitudinal vs. cross-sectional), the scores from the group of CI users assessed longitudinally appear to improve over time whereas the data from the control group (HA users) sampled cross-sectionally cannot. The use of single-point measures of hearing aid users' speech perception as controls for children with cochlear implants makes it difficult to interpret the speech perception increases found for cochlear implant users. That is, improvements in speech perception over time might also occur in children who use hearing aids simply as a result of maturation or additional aural rehabilitation.

In the second method described above (age-matched controls) the groups are often very small and there is a great deal of variability in speech perception scores for prelingually-deafened children. In the past, we have attempted to control for the effects of maturation by comparing the performance of children with CIs at one postimplant interval (e.g., 2 years postimplant) to that of hearing aid control subjects who were matched to CI users by age at onset of hearing loss and age at the time of testing (Kirk et al., 1995). However, because of the difficulty in finding matched-pairs of CI and HA users, this type of analysis can be applied to only limited numbers of subjects, and generalizations are difficult to make.

In the present study, we examined the speech perception measures collected from cochlear implant and hearing aid users over time in our laboratory. We used linear regression analyses of the speech perception scores from the HA users as a function of their age at testing. These linear regressions were used to generate predictions about the improvement in speech perception expected from HA users over time. Actual performance over time from CI users was then compared to the predicted improvements in performance for the unimplanted profoundly deaf HA users.

Because we have relatively large numbers of both Oral and Total Communication (TC, or the simultaneous use of signed and spoken English) HA users in our database, the linear regression analyses were performed on the data separately within each communication mode. The predictions of improvements in speech perception for the children using HAs for both modes of communication were compared to observed benefits children receive with cochlear implants. Such comparisons, as well as similar comparisons for improvements in speech production intelligibility and language between children with CIs and HAs (Svirsky, 1996), impact directly on the issue of cochlear implant candidacy. When a prelingually-deafened child is being considered for cochlear implantation, it is important that the clinicians working with the child have valid information about the expected gains in communication for that child based on that child's age, residual hearing, communication mode, and communication device.

The goals of the present study were: (1) to estimate the amount of improvement in speech perception scores that can be expected with maturation as a function of degree of hearing loss and mode of communication for the HA users; and (2) to compare the observed changes over time in speech perception by pediatric CI users to the improvements predicted for profoundly deaf HA users. Because the estimated improvements for the HA users may vary as a function of degree of hearing loss or communication mode, comparisons for the two modes of communication will be made independently (Oral-CI to Oral-HA; TC-CI to TC-HA). In carrying out these analyses, we assessed whether the observed speech perception skills of CI users met or exceeded the estimated skills of profoundly hearing-impaired children using HAs.

Methods

Participants

Participant characteristics for the 58 hearing aid users and the 74 cochlear implant users are presented in Tables 1, 2, and 3.

A. Hearing Aid Users

Fifty-eight (58) children with prelingual (congenital or onset at less than 3 years of age) profound hearing losses who used HAs participated in the study. The HA users were grouped in terms of their better ear unaided thresholds at 500, 1000, and 2000 Hz. Participants were identified as "Gold" HA users if two of the three threcholds were between 90 and 100 dB HL with none greater than 105 dB HL, and "Silver" HA users if two of the three thresholds were between 101 and 110 dB HL. Of the 24 children with the most residual hearing (i.e., Gold HA users), 15 used Oral Communication, and 9 used Total Communication. Of the 34 Silver HA users, 16 used Oral Communication and 18 used Total Communication (see Table 1). Thirteen of the Silver HA were subsequently implanted; thus the two participant groups (HA and CI) are not mutually exclusive. The cause of deafness for the majority of the children in the present study was unknown (66%). Meningitis was the cause of deafness for the highest percentage of children in which the cause of hearing loss was actually known (17%) (see Table 2). The average age at the onset of deafness and the mean age a hearing aid was fit were similar for the two groups of HA users, as well as for the two modes of communication (see Table 3).

Table 1.

Number of Participants by Communication Mode.

	Oral	TC	Total
CI	37	37	74
Gold HA	15	9	24
Silver HA	16	18	34

Table 2.

Number of Participants by Etiology of Hearing Loss.

Etiology	CI		HA			
			Gold		Silver	
	Oral	TC	Oral	TC	Oral	TC
Unknown	22	21	9	5	13	11
Meningitis	14	10	3	2	2	3
Mondini	0	1	0	0	0	0
Genetic	1	3	1	2	1	1
CMV	0	1	0	0	0	1
Viral Infection	0	0	1	0	0	0
Rubella	0	0	1	0	0	0
Febrile Seizures	0	0	0	0	0	2
Total	**37**	**37**	**15**	**9**	**16**	**18**

Table 3.

Age at Onset of Hearing Loss and Fitting with Device.

Devise		Onset (yrs)		Fit (yrs)	
		Oral	TC	Oral	TC
CI	**Mean**	0.35	0.64	5.72	4.78
	St. Dev.	0.73	0.82	1.58	1.72
	Range	0.0-3.0	0.0-2.8	2.7-8.9	2.2-8.7
Gold HA	**Mean**	0.71	0.41	2.29	1.29
	St. Dev.	0.90	0.88	1.26	0.75
	Range	0.0-3.0	0.0-2.5	0.8-5.0	0.3-2.8
Silver HA	**Mean**	0.05	0.31	1.56	1.23
	St. Dev.	0.14	0.62	0.88	0.83
	Range	0.0-0.4	0.0-2.0	0.3-4.1	0.1-3.0

B. *Cochlear Implant Users*

Seventy four (74) children with prelingual profound hearing loss who received the Nucleus 22-channel CI participated in the study. Of the 74 children with cochlear implants, 37 used Oral Communication and 37 used Total Communication (see Table 1). The cause of hearing loss for most of the CI users was unknown (61%). Just like the children wearing HAs, meningitis was the most common cause of hearing losses in which the etiology was known (32%) (see Table 2). The age at onset of deafness was quite similar for the CI users as compared to the HA users. The average age at implantation was similar for the children using Oral or Total Communication (see Table 3). Of the 47 children implanted for whom we have preoperative records, 32 were Bronze HA users (hearing loss > 110 dB HL at two of three frequencies, 500 Hz, 1000 Hz, or 2000 Hz), 14 were Silver HA users, and one was a Gold HA user.

CI Speech Processing Strategy

Children using CIs were tested with their current speech processor and coding strategy. In terms of the latest processing strategy used by the participants, five subjects used the F0F1F2 (Blamey, Dowell, Clark, & Seligman, 1987) strategy, 31 subjects used MPEAK (Skinner et al., 1991), and 38 subjects used SPEAK (Skinner et al., 1994). Five subjects had partial insertions. The number of active electrodes ranged from 8-22. As the technology of the CI speech processor and strategy improved during the course of the study, five subjects changed from the F0F1F2 to the MPEAK, and 18 subjects changed from the MPEAK to the SPEAK speech processing strategy. The number of subjects using each processor was similar for the Oral and TC groups. The new processors and processing strategies (SPEAK) should provide the user with more information and improved speech perception scores. However, if a CI user switches processors, it often takes a period of time for the user to become accustomed to the device, and scores on speech recognition tests can decrease during this period (Sehgal, Kirk, Svirsky, & Miyamoto, submitted). If all of the CI users had been using the most advanced speech processing strategy available at the present time for the Nucleus-22 device (SPEAK), the CI users may have achieved even higher speech perception scores than those seen in the present study.

Testing Intervals

Speech perception scores have been collected longitudinally from children with CIs for many years. Miyamoto, Osberger, Todd, Robbins, Karasek et al. (1994) compared longitudinal data from CI users to cross-sectional data from HA users. In the present study, we expanded on the Miyamoto et al. study in that we have attempted to collect and examine longitudinal data for both experimental (CI) and control (HA) groups. The data from neither the CI group nor the HA groups are strictly longitudinal. The vast majority of both the HA users and the CI users were tested more than once, but a few of the children were tested only once. In general, the CI users were tested at 6-month intervals if they were tested at Indiana University and yearly if they were implanted and tested elsewhere. Not every child was tested at each and every testing interval. Thus, the number of children tested at each time interval varied and was dependent on many factors including, but not limited to, distance from testing center and continued willingness to participate in the research study. Time is another factor in a longitudinal study; for the more recently-implanted children there are fewer data points. Thus, the data in the present study are not strictly longitudinal, nor can they be described simply as cross-sectional.

A. *Hearing Aid Users*

The Gold HA users were tested between one and six times at 6-month or yearly intervals. The Silver HA users were tested between one and four times at 6-month or yearly intervals. Thirteen of the Silver HA users were subsequently implanted.

B. Cochlear Implant Users

Participants were tested preoperatively as a baseline and then at 6-month or yearly intervals thereafter. Participants were tested for as long as 8.5 years postoperatively. Although rare, a few participants were tested only once postoperatively. Many of the children with cochlear implants were implanted at facilities other than Indiana University. These children are tested yearly for at least three years postoperatively as part of our longitudinal project.

Speech Materials

Participants were tested on a variety of speech perception materials in the Indiana University cochlear implant test battery. Two measures obtained from our test battery were selected for the present analysis (again, as an extension of Miyamoto, Osberger, Todd, Robbins, Karasek et al., 1994): the Minimal Pairs Test (Robbins, Renshaw, Miyamoto, Osberger, & Pope, 1988) and the Common Phrases Test (Osberger, Miyamoto et al., 1991). The Minimal Pairs test is an 80-item test designed to assess closed-set speech perception based on vowel and consonant feature recognition. One word (e.g., "bear") is presented to the child with two possible responses that differ by a single phoneme (e.g., "bear" and "pear"). Each pair of words appears twice, and each word from the pair acts as the stimulus once. Thus, if "bear" was the stimulus for the first presentation, "pear" would be the stimulus for the second presentation. Chance performance is 50%. Vowel height and place of articulation recognition are assessed with 32 of the 80 items, and consonant voicing, manner, and place of articulation are assessed with the remaining 48 items. The test is administered in an auditory-only mode. The children respond by identifying (or pointing to) a picture of the perceived stimulus.

The Common Phrases Test (Osberger, Miyamoto et al., 1991) assesses open-set sentence comprehension using three cue-presentation modalities: Auditory only (A), Visual only (V), and Auditory + Visual (A+V). The test items consist of six lists of ten two- to six-word phrases (simple commands or questions). The phrases are spoken to the child and the child responds by repeating the entire phrase or by correctly answering questions either orally or with the use of signed English. Scoring ranges from 0 - 100% in increments of 10% because only the number of completely correct responses is computed. The Visual-only modality is presented with the child's CI or HA turned off. We examined data from all three test conditions (A, V, A+V) in the present study. Both tests are administered via live-voice at approximately 70 dB SPL by an experienced audiologist or speech pathologist. When testing in the auditory modality, the experimenters covered their faces with an opaque mesh screen to eliminate visual cues.

Statistical Analysis

The data from the HA users were analyzed using linear regression analysis of speech perception scores as a function of age at testing to assess the effects of maturation (aging) on speech perception scores. First, linear regressions were carried out separately for the Oral and TC children in each of the hearing aid groups (Gold and Silver) for speech perception scores as a function of the age of the child at testing. This provides an estimate of the expected improvement in speech perception scores due to maturation in the absence of cochlear implantation. Second, the measures of speech perception by pediatric CI users collected over time were compared to the obtained regression lines, that is, to the estimates of performance for the hearing aid groups as a function of the age of the child at the time of testing. Again, comparisons were made separately within each communication modality.

Results

Minimal Pairs Test

Linear regressions (score vs. age at testing for the HA users) were performed for only the consonant subtest as the mean scores for the Vowel subtest approached the ceiling (100% correct) rapidly for all groups of children using either HAs or CIs and either Oral or Total Communication.

Insert Figure 1 about here.

The linear regression functions are shown in Figure 1. The top panel is for the Oral children and the bottom panel is for the TC children. Correlations are placed next to the corresponding line. Significant correlations are highlighted by asterisks (* $p < .05$, ** $p < .01$). Improvements in scores as a function of age at testing are predicted for both Gold and Silver HA users for both communication modes (positive slopes of the linear regression functions). Based on this analysis, consonant perception scores for the HA users would be expected to improve between 1 - 2% per year for a HA user irrespective of the mode of communication (Oral or TC) or the amount of residual hearing (Gold or Silver). As is shown in the top panel of Figure 1, for the children who use Oral communication, predicted performance for the Gold HA users is higher than predicted performance for the Silver HA users. The predicted scores for the Oral Gold HA users are approximately 20-30% higher at all points in time than the predicted scores for the Oral Silver HA users. This is to be expected because the Gold HA users have more residual hearing than the Silver HA users. As shown in the bottom panel of Figure 1, for the children who use Total communication, predicted performance is only slightly less than predicted performance of the children using Oral communication (top panel). There is a great deal of overlap in the predicted performance for the TC Gold and TC Silver HA users. Performance is predicted to improve over time for both groups at a rate of 1-2% per year.

When the data from the children using CIs are examined, the results follow the same general trend as the findings from Miyamoto, Osberger, Todd, Robbins, Karasek et al. (1994) in that the performance on the Minimal Pairs Test increases with the length of implant use. The data for the Oral children are shown in Figure 2 and the data from the children using TC are shown in Figure 3. The data derived from the linear regressions shown in Figure 1 for the HA users are plotted as points connected by lines and the data from the CI users are plotted as bars. The points represent the predicted scores for the HA users at the mean age of the CI users at that particular testing interval.

As illustrated in Figures 2 and 3, the number of children with CIs being tested at the different postoperative intervals varied. Therefore, the corresponding average age of the children tested at each postimplant interval does not increase linearly because (a) the number of children tested at each interval is different, and (b) the actual testing interval is somewhat variable. For example, a child could be tested between 10-14 months postimplant for the "1-year" testing interval. The average ages of the CI users at the preoperative testing interval were 6.1 years for the Oral subjects and 5.8 years for the TC subjects. At the 1-year postoperative testing interval, the average ages for the CI users were 6.8 years for the Oral subjects and 6.9 years for the TC subjects (instead of 7.1 years and 6.8 years, respectively). Therefore, the predicted scores for the children using HAs do not follow straight lines. Rather, they are the predicted scores for the HA users (from Figure 1) at the average age of the CI users at that particular testing interval.

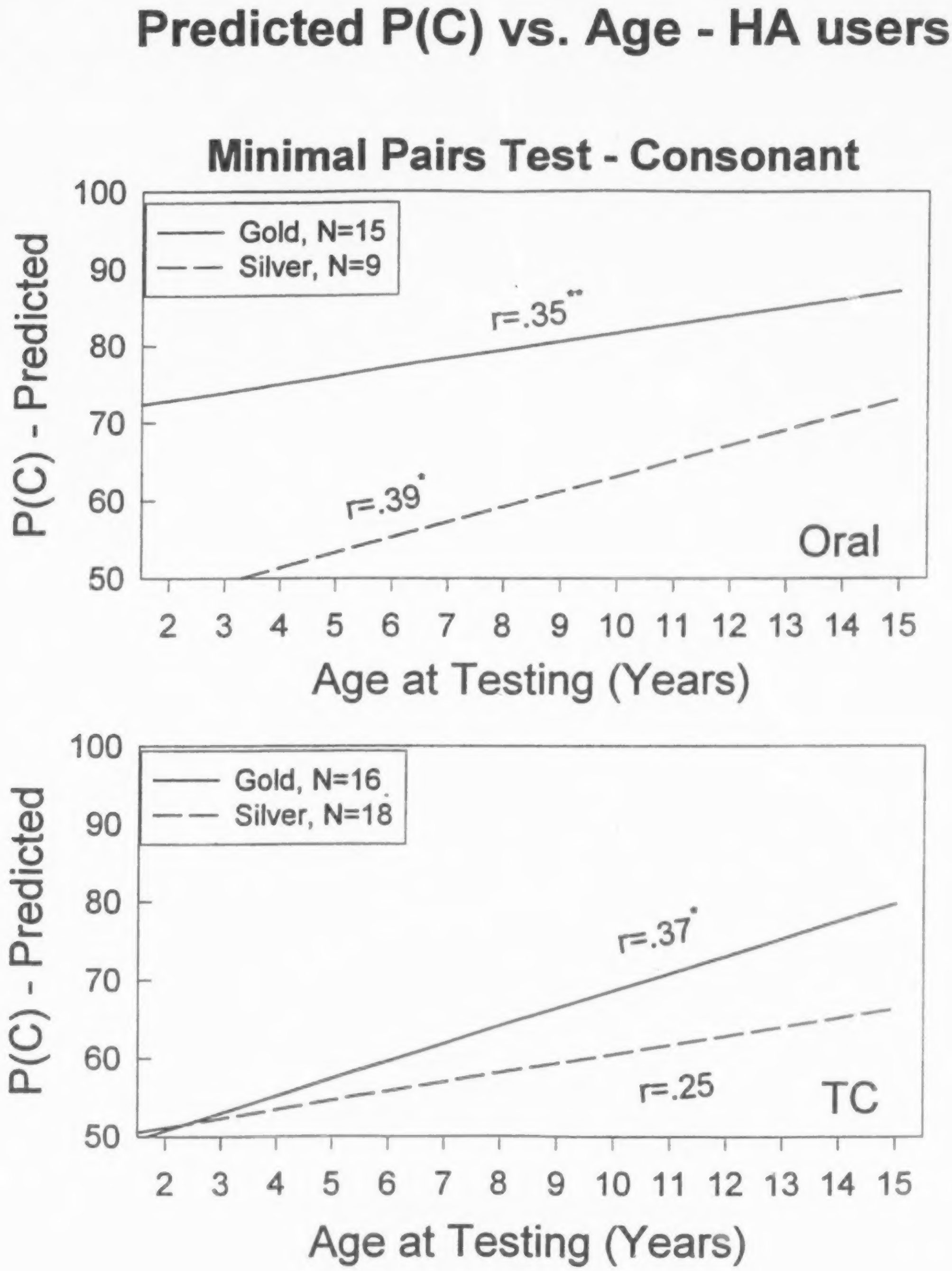


Figure 1. Linear regressions of test score vs. chronological age for the Minimal Pairs Test - consonant. Regression lines for the Oral children are in the top panel, and lines for the TC children are in the bottom panel. * $p < .05$, ** $p < .01$.

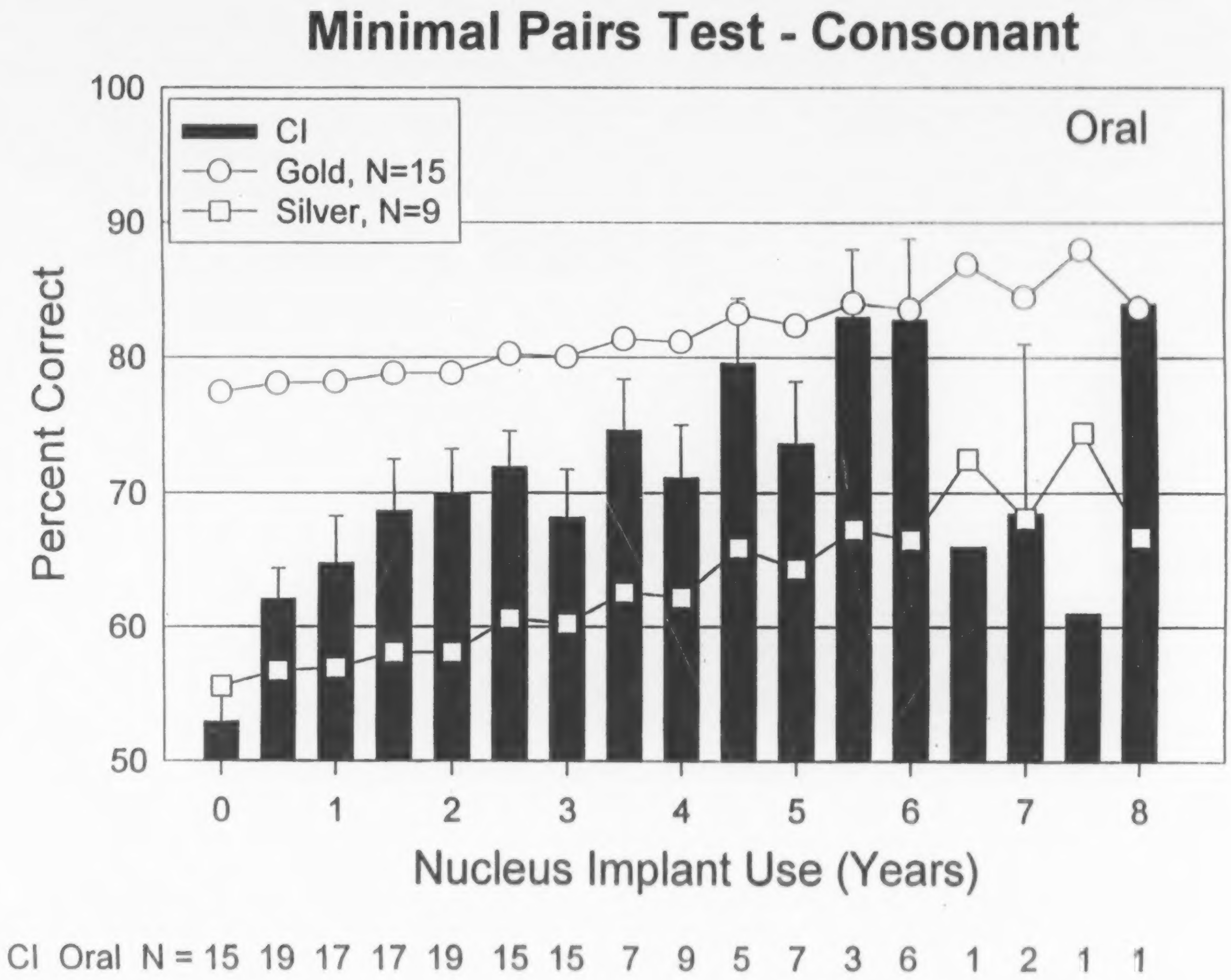


Figure 2. P(C) vs. implant use for the Minimal Pairs Test - consonant for Oral children. Data for the Gold and Silver HA users are predictions from the linear regressions in Figure 1 based on the average age of the CI users at a particular testing interval. Chance performance (50%) is shown by the dashed line. Error bars represent standard errors of the mean.

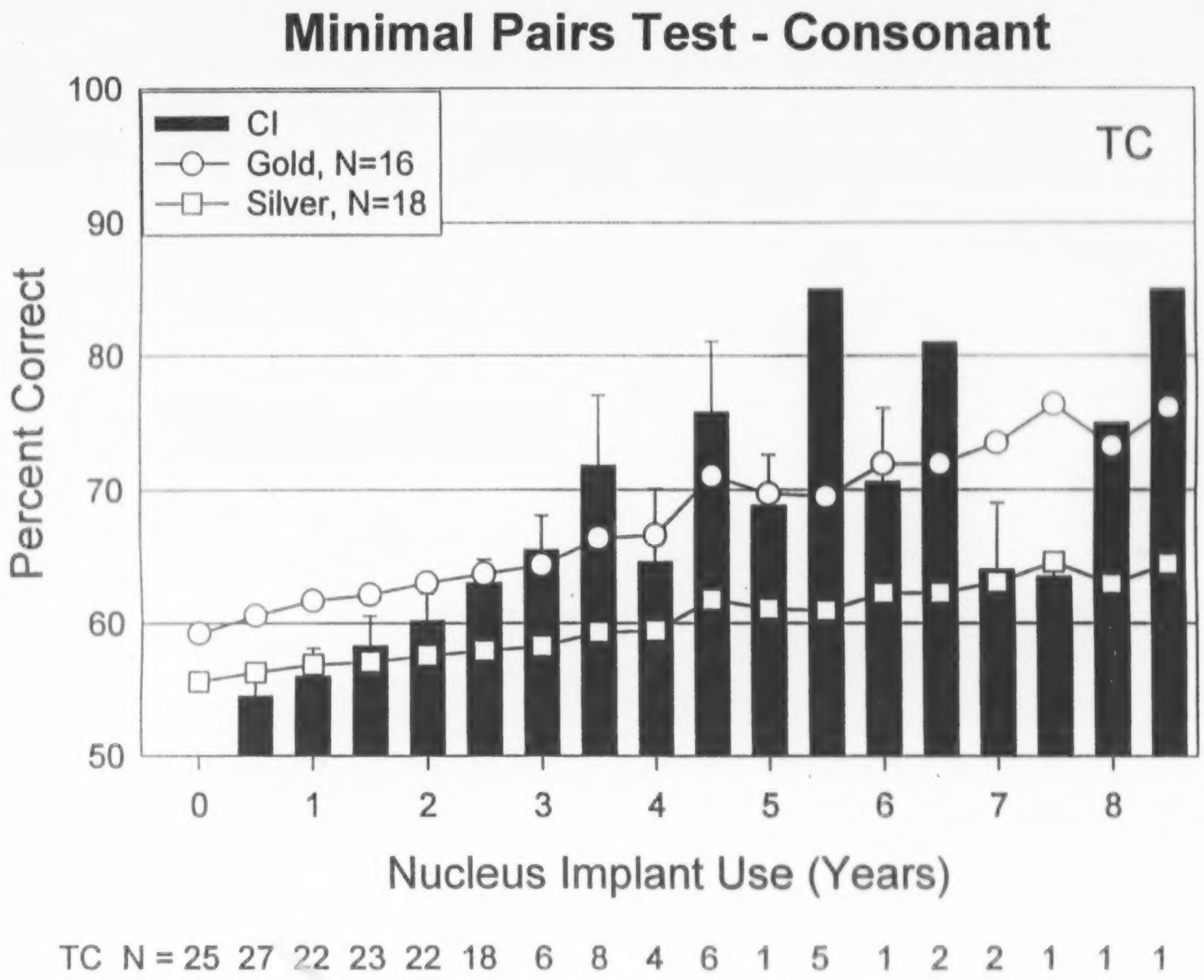


Figure 3. P(C) vs. implant use for the Minimal Pairs Test - consonant for the TC children. Data for the Gold and Silver HA users are predictions from the linear regressions in Figure 1 based on the average age of the CI users at a particular testing interval. Chance performance (50%) is shown by the dashed line. Error bars represent standard errors of the mean.

Insert Figures 2 and 3 about here.

Again, we are comparing speech perception scores for children using the same mode of communication. Chance performance on the Minimal Pairs Test is 50%, and a score of 65% on the 48-item consonant perception subtest is considered significantly greater than chance (p=.03). Scores for both Oral and TC CI users were at chance levels (50%) just prior to implant. For the Oral children, the mean consonant perception score reached a level statistically greater than chance (65%) by approximately one year postimplantation, and the mean score of the TC users reached this level (65%) by approximately 3 years after implantation. For the Oral children, the average score for the CI users approached the predicted score for the Gold HA users after approximately 5 years of implant use. For the children using TC, the mean scores for the CI users were approximately equal to the predicted scores for the Gold HA users after approximately 3 years of implant use. The average scores for the Oral children using CIs were slightly higher on the Minimal Pairs Test than the average scores for the children using TC. The number of participants tested with long (> 5 years) implant use is quite small, and the associated data should be interpreted with some caution.

Common Phrases Test

A. *Hearing Aid Users*

Separate linear regressions were performed (score vs. age at testing for the HA users) for the A, V, and A+V test conditions for the TC and Oral children. The linear regression functions are shown in Figure 4.

Insert Figure 4 about here.

The data for the Oral children are shown in the left two panels (Gold - top, Silver - bottom), and the data for the children using TC are shown in the right two panels (Gold - top, Silver - bottom). Correlations are placed next to the corresponding line. Significant correlations are highlighted with asterisks (* $p < .05$, ** $p < .01$). Improvements in scores as the age at testing increases (positive slopes of approximately 3% increase in score per year) are predicted for the Oral Gold and Oral Silver HA users for the A (audition alone) condition. As with the Minimal Pairs Test, the predicted scores for the Oral Gold HA users (left panels) were substantially higher (they obtained 25-40% higher scores under the A condition) than scores for the Oral Silver HA users.

For the participants using TC (right panels), no improvement in speech perception as a function of age is predicted for the A condition (slope of approximately 0% increase in score per year). The average score for the TC Gold HA (35%) users is substantially higher than the average score for the TC Silver HA users (6%).

The results from the Common Phrases Test carried out in the V condition demonstrate that the rate of improvement over time was much greater for the Oral HA users (5.7 to 7.7 % improvement per year) than for the TC HA users (-1.0 to 2.0 % improvement per year). The predicted scores for the TC Gold HA users were higher than the predicted scores for the TC Silver HA users. This difference is related to the negative slope (although not significantly different from 0.0) of the regression line for the Silver HA users.

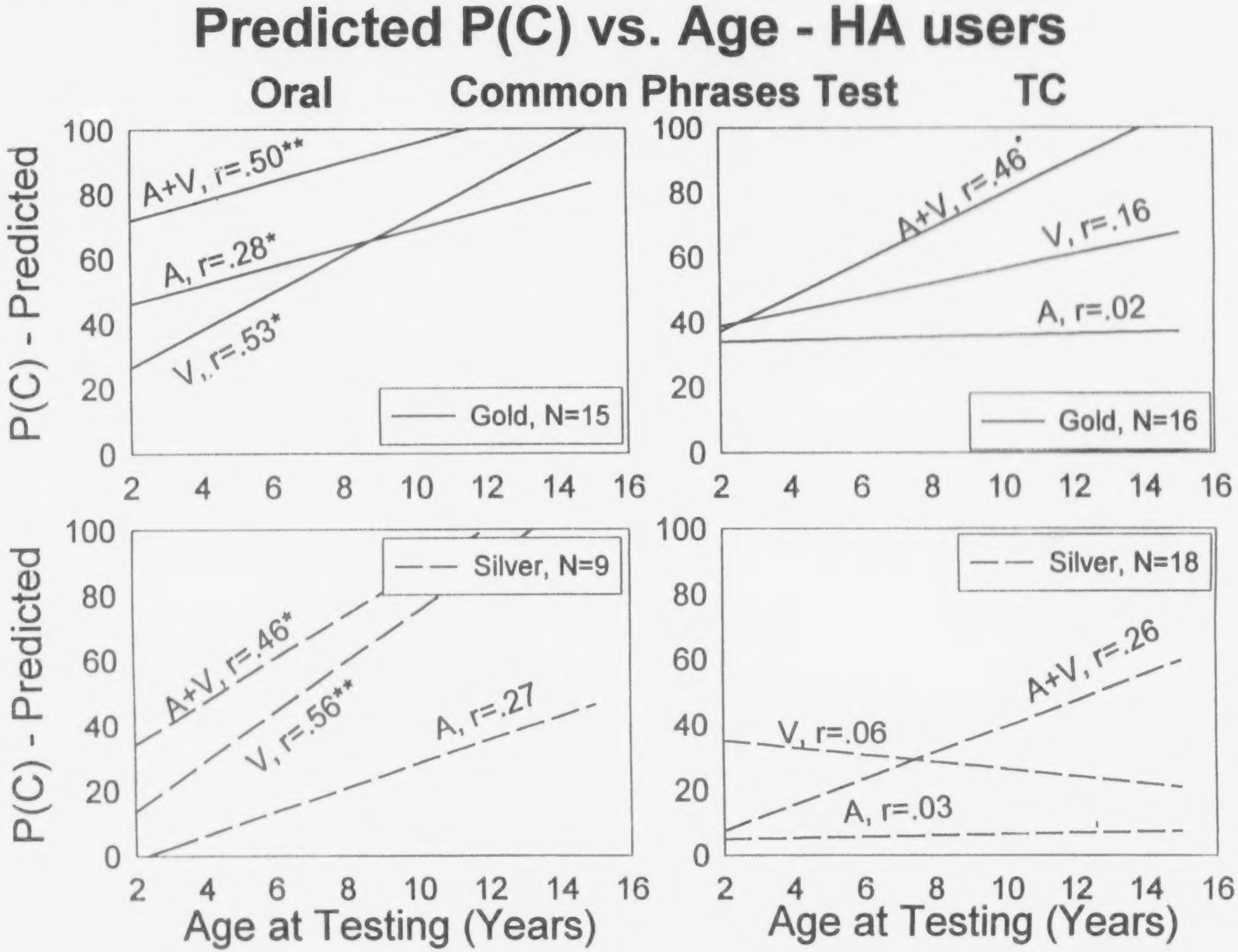


Figure 4. Linear regressions of test score vs. chronological age for the Common Phrases Test. Regression lines for the Oral children are in the left two panels (Gold - top, Silver - bottom), and lines for the TC children are in the right two panels. Regression lines were generated for the Audition alone (A), Vision alone (V), and audition plus vision (A+V) conditions. * p < .05, ** p < .01.

For the Oral children, the regression lines for both the Gold and Silver HA users overlap, and differences between the two groups are small (±10%).

When visual cues were added to the auditory condition, large rates of improvement with age at testing were predicted (3-6% improvement per year) for both Silver and Gold HA users and both Oral and Total Communication modes. Predicted performance for the Gold HA users was higher than predicted performance for the Silver HA users for both communication modes. For the Oral children (left panels), the regression lines predict that the difference between the Gold and Silver HA users decreases over time, and by approximately age 12, the difference in A+V speech perception between the two groups is negligible, and that performance for both groups should be approximately 100%. For the TC users (right panels), the regression lines predict that the Gold HA users obtain 25-50% more open-set sentence recognition than the Silver HA users. As expected, predicted performance for the A+V condition is greater than performance predicted for either the A or the V conditions alone.

B. Cochlear Implant Users

Of the 73 children with CIs tested with the Common Phrases Test, 59 (80.8%) acquired at least some [P(C) > 0%] open-set sentence recognition under the auditory-only (A) condition at some point during testing. Of the 37 CI participants in oral communication programs, 35 achieved some open-set speech recognition during testing. Of the 36 CI participants in total communication programs, 24 achieved some open-set speech recognition during testing. The mean scores on the Common Phrases Test for the CI users at the different testing intervals are shown in Figures 5 - 8.

Insert Figures 5, 6, 7 and 8 about here.

As shown in Figures 5 through 8, performance on the Common Phrases Test increases over time for the CI users for the audition alone (A), vision alone (V), and the audition plus vision (A+V) conditions. Data from the Oral children are shown in Figures 5 and 6, and data from the children using TC are shown in Figures 7 and 8. Overall, the Oral children using CIs perform better than the TC children using CIs on the Common Phrases Test under all three conditions, the A, V, and the A+V conditions. The same result (Oral scores greater than TC scores) was seen in the predicted regression lines for the children using HAs in Figure 4.

1. Oral When the data from the Oral children in the audition alone (A) condition were analyzed (Figure 5 - right panel), the average score for the CI users at the pre-implant interval (10%) was below that predicted for the Silver HA users (15%). By one year of implant use, the average score for the CI users increased to a level (40%) that was much greater than that predicted for the Silver HA user (18%). The amount of benefit the children using CIs derive from their devices continues to be greater than that predicted for the Silver HA users over time, but scores on the Common Phrases Test under the A condition for the CI users remain approximately 20% less than the predicted scores for the Gold HA users. The predicted scores from the children in both HA groups improve at approximately the same rate (3-4% improvement per year). The estimated rate of improvement for the CI users is approximately 7% per year.

In the V condition (Figure 5 - left panel), the absolute scores as well as the rates of improvement are similar for both groups of HA users as well as the CI users. Scores increased from approximately 50% at one year post-implant to 70% at five years post-implant.

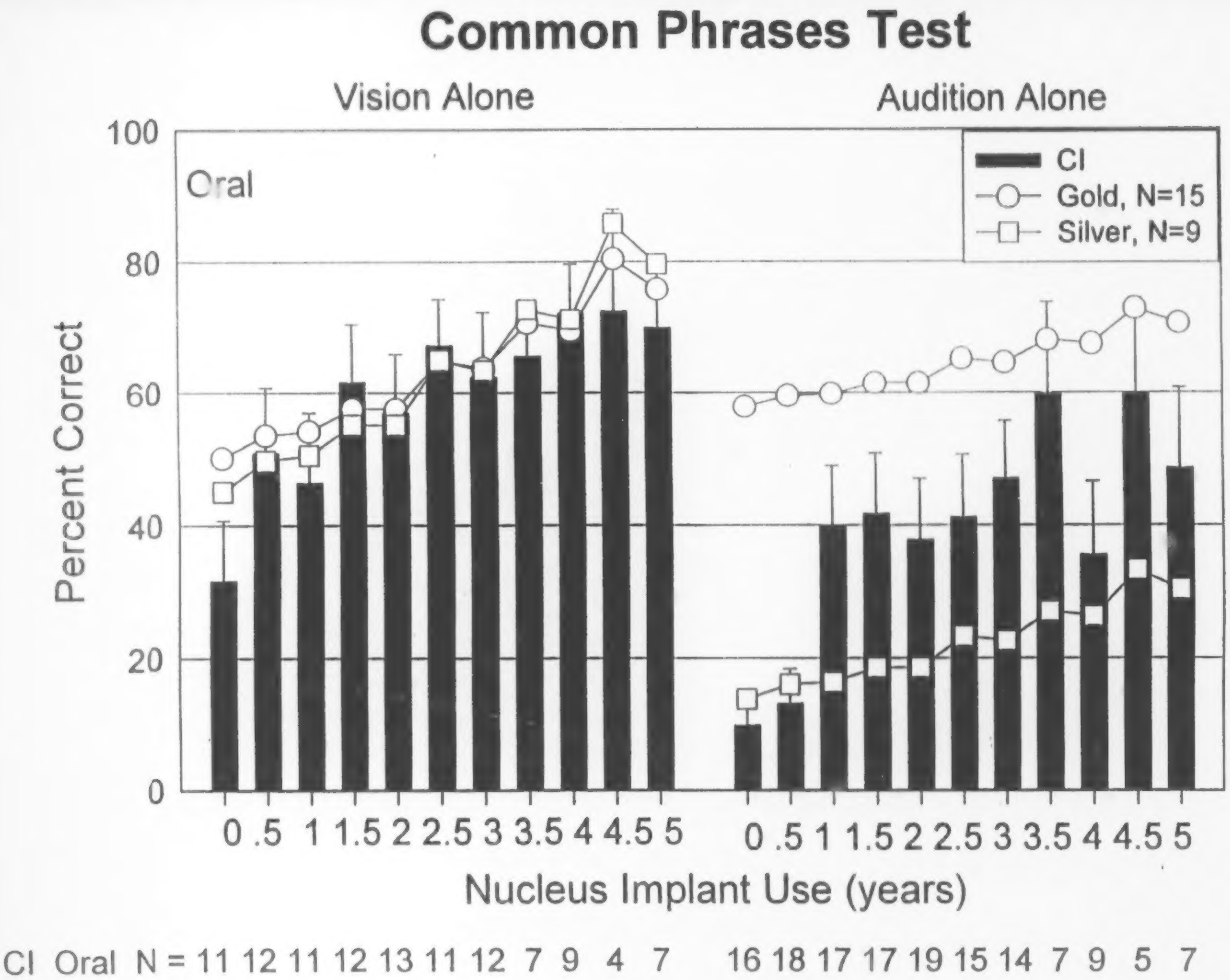


Figure 5. P(C) vs. implant use for the Common Phrases Test - Oral children. The vision alone (V) condition is plotted on the left side of the figure. The audition alone (A) condition is plotted on the right side of the figure. Data for the Gold and Silver HA users are predictions from the linear regressions in Figure 4 based on the average age of the CI users at a particular testing interval. Error bars represent standard errors of the mean.

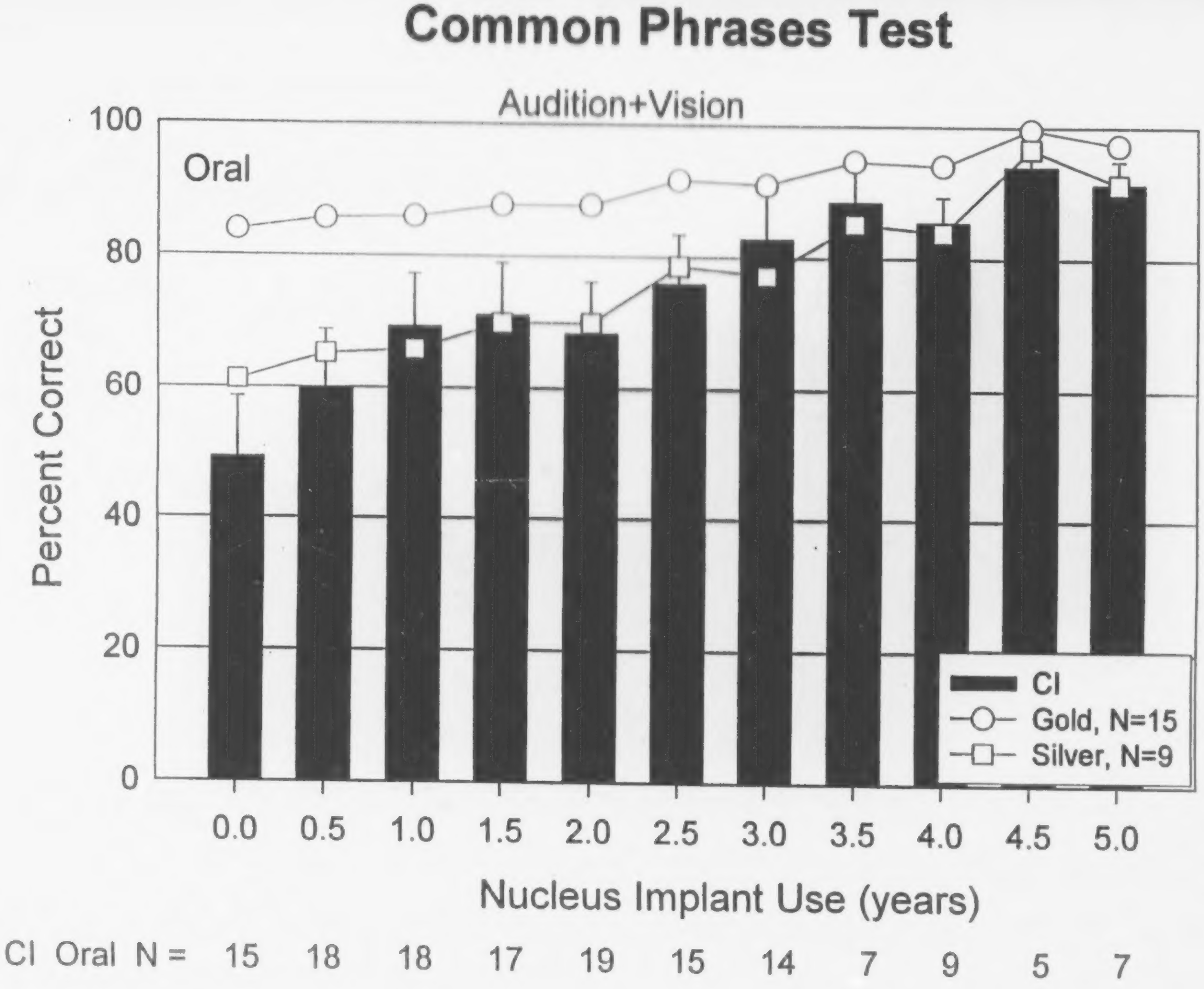


Figure 6. P(C) vs. implant use for the Common Phrases Test - Oral children in the audition plus vision (A+V) condition. Data for the Gold and Silver HA users are predictions from the linear regressions in Figure 4 based on the average age of the CI users at a particular testing interval. Error bars represent standard errors of the mean.

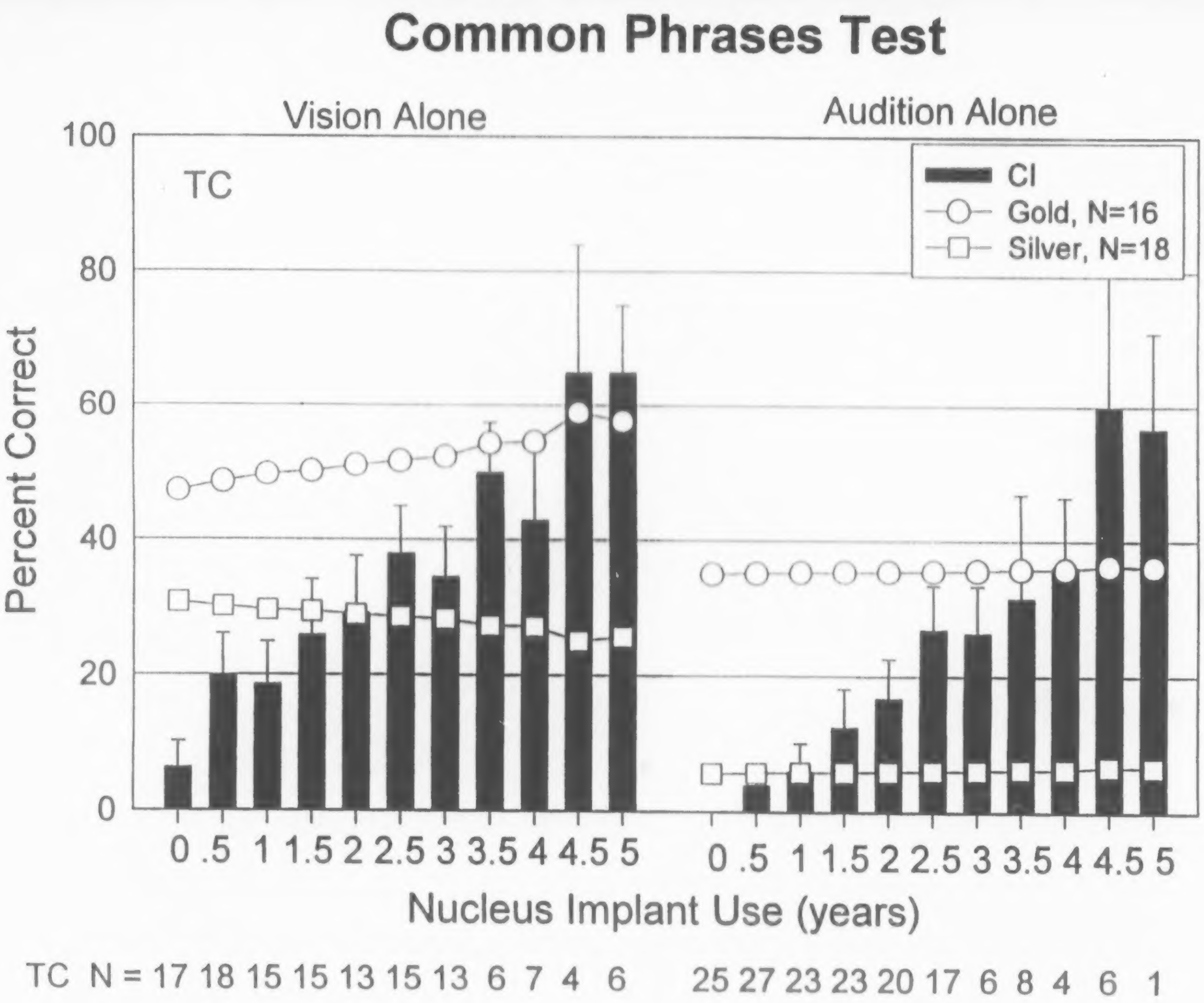


Figure 7. P(C) vs. implant use for the Common Phrases Test - TC children in the A and V conditions. The vision alone (V) condition is plotted on the left side of the figure. The audition alone (A) condition is plotted on the right side of the figure. The data for the Gold and Silver HA users are predictions from the linear regressions in Figure 4 based on the average age of the CI users at a particular testing interval. Error bars represent standard errors of the mean.

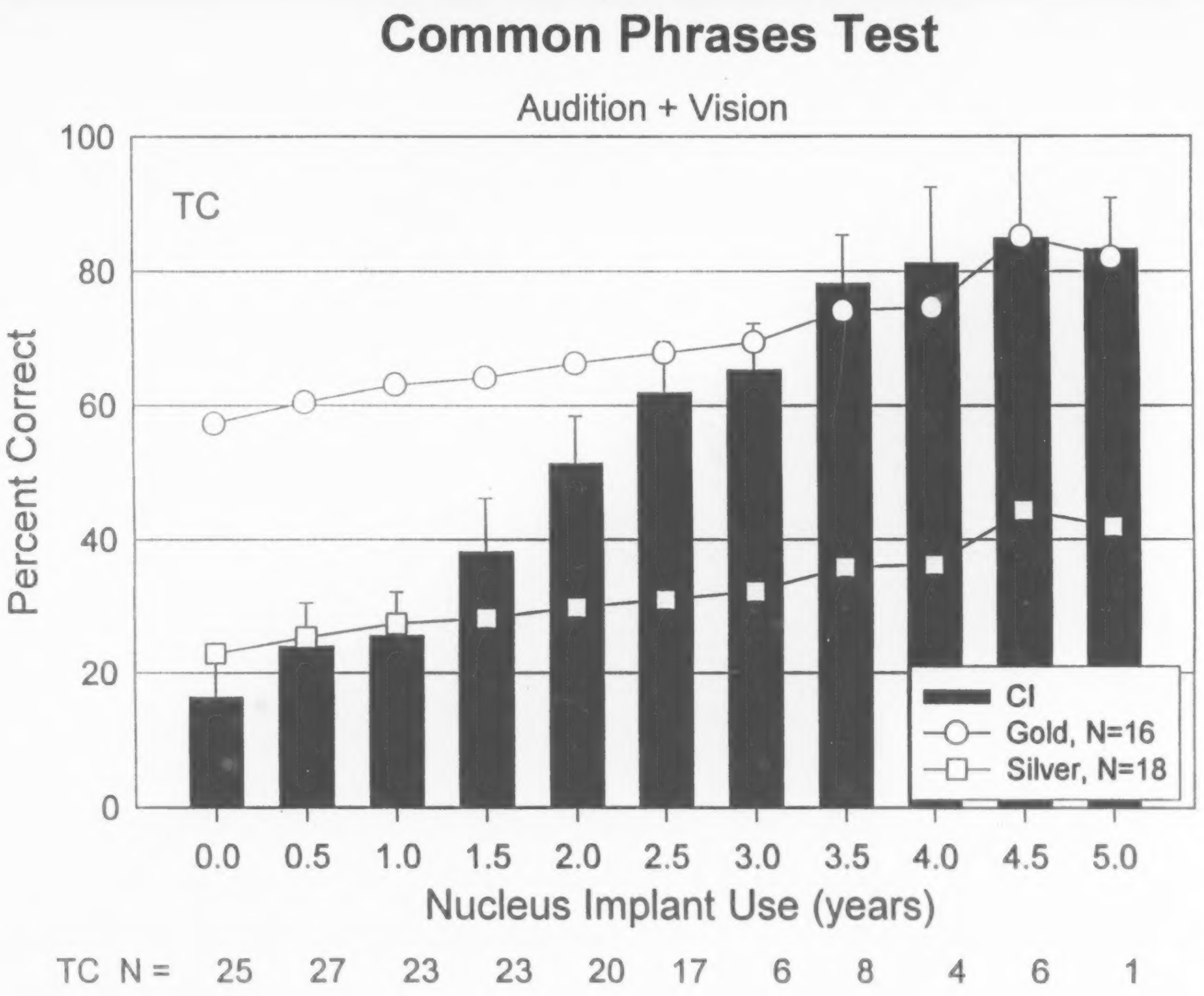


Figure 8. P(C) vs. implant use for the Common Phrases Test - TC children in the A + V condition. Data for the Gold and Silver HA users are predictions from the linear regressions in Figure 4 based on the average age of the CI users at a particular testing interval. Error bars represent standard errors of the mean.

In the A+V condition (Figure 6), the mean score for the CI users is approximately 50% at the preimplant interval, which is approximately 12% less than the predicted score for the Silver HA users (62%), and about 35% below the predicted score for the Gold HA users (85%). By one year postimplant, the mean score for the CI users has attained the level predicted for the Silver HA users, and by five years post-implant, the mean score for the CI users is approximately 90% and essentially equivalent to the predicted score for the Gold HA users. The rate of improvement observed for the CI users was similar to the rate of improvement predicted for the Silver HA users, and greater than that predicted for the Gold HA users. The relation between the scores for the different groups may be confounded by a ceiling effect for this particular test. The addition of visual cues to the auditory input improves performance on the Common Phrases Test a great deal, and after several years of device use (HA or CI), the children are, in general, performing very well on the test. Differences in performance between groups are difficult to assess under the A+V condition when average performance is nearly perfect.

2. TC When the data from the children using TC in the A condition are analyzed (Figure 7 - right panel), for both the Gold and Silver HA users, the performance under the A condition is not correlated with the age at testing (r 0.0). The best estimate of performance is the mean of the test scores under this condition. The predicted (average) score for the Gold HA users is approximately 35%, and the average score for the Silver HA users is approximately 6%. Mean scores for the CI users surpass the average score for the Silver HA users after 1.5 years of use, and they reach the average score for the Gold HA group (35%) by approximately four years of implant use. The estimated rate of improvement seen in the TC CI users is high (12% improvement per year).

In the V condition (Figure 7 - left panel), scores for the CI users reach levels predicted for the Silver HA users by approximately 2 years of implant use and levels predicted for the Gold HA users after about 5 years of implant use. Scores for the TC CI users also improve at a fast rate (11% improvement per year).

When both auditory and visual cues (A+V) are available, the predicted scores for the HA users increase with age (Figure 8). The absolute scores for the Gold HA users are approximately 35% higher than scores for the Silver HA users. Scores for the children using CIs surpass the predicted scores for the Silver HA group after approximately 2 years of use, and they reach the levels of the Gold HA users by approximately 4 years of implant use. The predicted rate of improvement is similar for both Silver and Gold HA users (4-5% improvement per year), and the estimated rate of improvement for the CI users is greater (15% improvement per year) than that for either the Gold or the Silver HA users.

Discussion

In this study, we examined the relation between speech perception scores (Minimal Pairs Test - Closed Set, Common Phrases Test - Open Set) and age at testing for prelingually-deafened children who use hearing aids to estimate the increase in performance due to maturation in the absence of cochlear implantation. We grouped the children based on their mode of communication, Oral or TC, and computed regression equations. These predictions were compared to data obtained from prelingually-deafened children with cochlear implants. On the Minimal Pairs Test, the children with CIs obtained at least as much consonant feature recognition as the Silver HA users. On the Common Phrases Test, the scores from the children with CIs approached the scores predicted for the Gold HA users.

The findings from the present study and results from other research with prelingually-deafened children impact directly on the management of deaf children. We are trying to answer questions about the expected speech perception benefits of hearing aids and cochlear implants to deaf children in different educational settings. Our results suggest that, on average, children with hearing losses in the 101-110 dB HL range (Silver HA users) would receive greater speech perception benefits from a CI than they do from their hearing aids irrespective of the mode of communication they are currently using. This result lends further support to the earlier study by Miyamoto, Osberger, Todd, Robbins, Karasek et al. (1994). They found that the speech perception scores of their implant users increased over time to be greater than the mean scores of their Silver HA users obtained at a single point in time. The present results demonstrate that even though the speech perception skills of Silver HA users increase over time, they do not keep pace with the average gains achieved by children who receive a cochlear implant. Similar results were also reported in a recent study by Geers and Brenner (1994) in which the speech perception of their CI users was similar to the speech perception of HA users with losses in the 90-100 dB HL range, and it was better than the speech perception of their HA users with losses greater than 100 dB HL.

Research concerning the speech perception skills of prelingually-deafened children also impacts directly on the issue of implantation candidacy. Decisions as to the audiological criteria used to determine which children should even be considered for a cochlear implant arise from this line of research. The current recommendations from the most recent NIH Consensus Conference on Cochlear Implants (NIH Consensus Conference, 1995) suggest that children with profound (>90 dB HL) bilateral sensorineural hearing loss and minimal speech perception may be considered for cochlear implantation. In a recent study, Zwolen et al. (1997) examined speech perception skills in two groups of prelingually-deafened children who received cochlear implants. The first group fit the recommended audiological criteria and had no open-set speech recognition prior to implantation. The second group also fit the recommended audiological criteria, but they had some open-set speech recognition prior to implantation. Both groups (including the children who had some open-set speech recognition prior to implantation) received a great deal more benefit from their implants than they did with their hearing aids. Although the authors did not attempt to estimate how these children would have performed if they had continued with hearing aids, they suggested that the child may benefit more from a CI than from continued HA use. Zwolen et al. further suggested that the selection criteria for CI candidacy should be broadened to include children who receive some open-set speech recognition.

The results from the present study support the suggestion of Zwolen et al. (1997) that the audiometric criteria for cochlear implantation might be broadened as our data indicate that CI users' speech perception skills exceed those of Silver HA users, and approach the skills of Gold HA users. However, although audiological criteria are important in the decision process for implantation, they are only a small part of the CI candidacy determination. CI candidacy is not an issue that should be taken lightly, and children should continue to be evaluated for a CI on an individual basis. Clinicians and parents of deaf children need up-to-date information as to the options available to them and the gains in communication skills they should expect to see for the children at different ages with the different devices in different educational settings. Without current data, it is even more difficult for the parents of deaf children to make the decisions that will impact the lives of their children.

The data clinicians need to help parents make decisions about implantation are often difficult to collect, report, and interpret. First of all, for ethical as well as logistic reasons, it is impossible to carry out a typical double-blind, randomized treatment vs. control study. Second, if children with CIs are the treatment group, it is not clear which children should be the control group. Should profoundly deaf children who use hearing aids, tactile aids, or no sensory aids act as a control group for children with CI? We chose children who use hearing aids as a control group for several reasons. Implant candidacy requires the child to participate in a hearing aid trial under close observation often for several months. The child's performance with a hearing aid is closely monitored, and performance with the hearing aids must be less than the average performance of children with a cochlear implant. The children showing little-to-no gain with hearing aids are given the option of cochlear implantation. The children who demonstrate adequate gains in speech perception with the hearing aids (usually those children with more residual hearing, i.e., Gold HA users) do not, in general, receive a cochlear implant. Third, the logistics of these kinds of clinical studies make data collection difficult, and the reporting and interpretation of the results even more difficult. Typically, subject attrition occurs over time as families move, children stop using their implants or hearing aids, or patients choose to return for visits less over time. To demonstrate this point, a recent study from the University of Iowa (Fryauf-Bertschy et al., 1997) found that 14 of the 40 children (35%) who received a cochlear implant at their institution (all used Total Communication) were minimal or non-users of their devices. The children described as "minimal" or non-users tended to be older at the time of implantation, and performed at lower levels on speech perception tasks than the full-time implant users. It is difficult to determine from the University of Iowa data, however, whether lower scores on speech recognition tasks are directly attributable to the age at implantation, the lack of device use, or some other variable or combination of variables. It is also difficult to determine whether the lack of device use is a behavioral response to lack of usable auditory information, peer pressure, or another reason.

In the same light, it is also possible that families of children who perform better with a device may opt to return for more testing sessions than families with children who do not perform well or are unhappy with their device, thus biasing the study. Finally, many clinical studies, such as those involving cochlear implants, are dependent upon the technological advances arising in the field during the time period of the study. As newer speech processors and coding strategies are developed, device manufactures want to offer their clients an improved product, and patients often change devices in hopes of receiving more clinical benefit. For these, and numerous other reasons, clinical comparisons across time for cochlear implant users are challenging. These difficulties exist at our institution and probably exist at all other institutions where clinical research on the benefit of cochlear implants is conducted.

Nevertheless, we have attempted to overcome some of the problems mentioned above by collecting and examining some longitudinal data from a control group of children using HAs as well as our treatment group of children with CIs. As the data from the HA users (or CI users) were not strictly longitudinal, we used linear regression analysis to predict improvements in speech recognition scores as a function of age or maturation for the children using HAs (control group). These predictions were compared to observed data from children with CIs to determine if use of a CI helps a child attain better speech perception scores than would be expected if the child were not implanted and used high-powered hearing aids instead. Using linear regression techniques is an improvement over past comparisons between the two subject groups (Miyamoto, Osberger, Todd, Robbins, Karasek et al., 1994), but more longitudinal data from children (both HA users and CI users) is needed to allow us to make better predictions about the potential benefit to speech perception from implantation.

We did not specifically test the effect of communication mode on speech perception scores. If a test of the effect of communication mode on test score would prove to be significant, one must be careful not to generalize the result to infer that a particular mode of communication is superior to another. It may be that children in Oral programs have more residual hearing than the children in TC programs and receive more auditory information through their hearing aids. For the children who receive little-to-no useful auditory information with a hearing aid and are implanted, the children in oral programs are more dependent upon the information they receive from the CI for communication than are the children in TC programs who can use and rely upon sign language for communication. It is also possible that the children in oral programs receive more training focused on the development of particular auditory and speech skill assessed by the speech perception testing protocols than the children in TC programs (Quittner & Steck, 1991).

Many factors are involved in the decision about which communication mode a child will use. One of the most important factors in deciding upon a communication mode for a prelingually-deafened child is the willingness of the family to learn and use the chosen communication mode. The availability of education and rehabilitation services for the child are also important in deciding upon a mode of communication. If the child is not able to make adequate gains in language development or academic progress with a chosen mode of communication, then the family should consider a different communication mode for the child.

Summary and Conclusions

We examined speech perception scores using two measures, the Minimal Pairs Test, a closed-set speech discrimination test based on consonant feature perception, and the Common Phrases Test, an open-set sentence comprehension test, over time for prelingually-deafened children using CIs or HAs and either Oral or Total Communication. In general, scores for CI users were at least as high as scores for Silver HA users and approached scores for the Gold HA users after several years of device use.

On the Minimal Pairs Test, for the children using CIs, scores for both Oral and TC children improved at a rate comparable to the rate of improvement seen in the profoundly hearing-impaired children using HAs with the most residual hearing ("Gold") using the same mode of communication. Scores from both Oral and TC users are similar - Oral children performed only slightly better on the Minimal Pairs Test than did the children using TC. This is to be expected because the Minimal Pairs Test assesses speech discrimination based on single-feature minimal pairs in a two-alternative, forced-choice design (chance performance is 50%). Furthermore, the closed-set scores for the children with CIs are similar to and improve at a rate comparable to the rate of improvement seen in the Gold HA users. The number of subjects tested with long implant use (> 5 years) is quite small, and the associated data should be interpreted with some caution.

On the Common Phrases Test, the scores for the TC children with CIs approach and probably surpass scores for the Gold HA users in the A and A+V conditions after approximately 4 years of device use. The rate of improvement in scores over time tor the TC children with CIs is greater than the rate of improvement predicted for the TC children using Gold HAs. The scores for Oral children with CIs improve at a rate greater than that predicted for the Oral Silver HA users, but their scores do not reach the levels of scores for the Oral Gold HA users after five years of implant use. The actual scores for the Oral children with CIs fall in between the scores for the Silver and Gold HA users in the Common Phrases Test in the A condition, and they are similar to the scores from the Silver HA users when visual cues are added (A+V).

Results from studies of speech perception, speech production, and language development are all needed to help clinicians determine the expected overall benefits to prelingually-deafened children from cochlear implantation versus hearing aid or other sensory aid. Speech intelligibility and language development have been examined by Svirsky (1996). His results, from an analysis similar to the one used in the present study, demonstrate that the intelligibility of the speech of children with CIs reaches the levels of intelligibility of the Silver HA users after 1-2.5 years of use, and moreover, the language scores of children with CIs improve faster than the language scores for the Gold HA users. It is apparent from the present study that prelingually-deafened children obtain higher speech perception scores with a cochlear implant than they do with a hearing aid if the amount of residual hearing is in the Silver range (101-110 dB HL). Thus, in terms of improvements in overall communication skills, our data suggest that prelingually-deafened children with enough residual hearing to be classified as Silver HA users would benefit more from a CI than a conventional HA, and at least some of the children who would be Gold HA users might benefit more from a CI than from a HA. As the CI technology improves, newer implants and speech processors will help CI users achieve even greater communication benefit than they do now. It remains to be seen whether children with slightly more residual hearing [Gold HA users (90-100 dB HL), or even children with severely impaired hearing (70-90 dB HL)] will achieve greater overall levels of communication with a hearing aid or a cochlear implant.

References

Blamey, P. J., Dowell, R. C., Clark, G. M., & Seligman, P. M. (1987). Acoustic parameters measured by a formant-estimating speech processor for a multiple-channel cochlear implant. *Journal of the Acoustical Society of America, 82,* 38-47.

Carney, A. E., Osberger, M. J., Miyamoto, R. T., Karasek, A., Dettman, D. L., & Johnson, D. L. (1991). Speech perception along a continuum: From hearing aids to cochlear implants. In Feigin, J. A. & Stelmachowicz, P. G. (Eds.) *Pediatric Amplification.* (pp. 93-113). Omaha, NE: Boystown National Research Hospital.

Cowan, R. S. C., Brown, C., Whitford, L. A., Galvin, K. L., Sarant, J. Z., Barker, E. J., Shaw, S., King, A., Skok, M., Seligman, P. M., Dowell, R. C., Everingham, C., Gibson. W. P. R., & Clark, G. M. (1994). Speech perception in children using the advanced SPEAK speech-processing strategy. *Annals of Otology, Rhinology, and Laryngology, 104* (Suppl. 166), 318-321.

Fryauf-Bertschy, H., Tyler, R. S., Kelsay, D. M., & Gantz, B. J. (1992). Performance over time of congenitally deaf and postlingually deafened children using a multichannel cochlear implant. *Journal of Speech and Hearing Research, 35,* 913-920.

Fryauf-Bertschy, H., Tyler, R. S., Kelsay, D. M.,Gantz, B. J., & Woodworth, G. G. (1997). Cochlear implant use by prelingually deafened children: The influences of age at implant and length of device use. *Journal of Speech, Language, and Hearing Research, 40,* 183-199.

Gantz, B. J., Tyler, R. S., Tye-Murray, N., & Fryauf-Bertschy, H. (1994). Long term results of multichannel cochlear implants in congenitally deaf children. In Hochmair-Desoyer, I. J. & Hochmair, E. S. (Eds.) *Advances in Cochlear Implants* (pp. 528-533). Vienna, Austria: Manz.

Gantz, B. J., Tyler, R. S., Woodworth, G. G., Tye-Murray, N., & Fryauf-Bertschy, H. (1994). Results of multichannel cochlear implants in congenital and acquired prelingual deafness in children: Five-year follow-up. *American Journal of Otology, 12* (Suppl.), 1-7.

Geers, A. & Brenner, C. (1994). Speech perception results: Audition and lipreading results. *Volta Review, 96,* 97-108.

Kirk, K. I., Osberger, M. J., Robbins, A. M., Riley, A. I., Todd, S. L., & Miyamoto, R. T. (1995). Performance of children with cochlear implants, tactile aids, and hearing aids. *Seminars in Hearing, 16,* 370-381.

Miyamoto, R. T., Kirk, K. I., Robbins, A. M., Todd, S., Riley, A. (1996). Speech perception and speech production skills of children with multichannel cochlear implants. *Acta Otolaryngologica (Stockholm), 116,* 240-243.

Miyamoto, R. T., Kirk, K. I.,Todd, S. L., Robbins, A. M., & Osberger, M. J. (1995). Speech perception skills of children with multichannel cochlear implants or hearing aids. *Annals of Otology, Rhinology, and Laryngology, 104* (Suppl. 166), 334-337.

Miyamoto, R. T., Osberger, M. J., Robbins, A. M., Myres, W. A., & Kessler, K. (1993). Prelingually deafened children's performance with the Nucleus multichannel cochlear implant. *American Journal of Otology, 14,* 437-445.

Miyamoto, R. T., Osberger, M. J., Robbins, A. M., Myres, W. A., Kessler, K., & Pope, M. L. (1991). Comparison of speech perception abilities in deaf children with hearing aids or cochlear implants. *Otolaryngology- Head & Neck Surgery, 104,* 42-46.

Miyamoto, R. T., Osberger, M. J., Todd, S. L., Robbins, A. M., Karasek, A., Dettman, D., Justice, N., & Johnson, D. (1994). Speech perception skills of children with multichannel cochlear implants. In Hochmair-Desoyer, I. J. & Hochmair, E. S. (Eds.) *Advances in Cochlear Implants* (pp. 498-502). Vienna, Austria: Manz.

Miyamoto, R. T., Osberger, M. J., Todd, S. L., Robbins, A. M., Stroer, M. A., Zimmerman-Phillips, S., & Carney, A. E. (1994). Variables affecting implant performance in children. *Laryngoscope, 104,* 1120-1124.

NIH Consensus Conference. (1995). Cochlear implants in adults and children. *Journal of the American Medical Association, 274,* 1955-1961.

Osberger, M. J., Miyamoto, R. T., Zimmerman-Phillips, S., Kemink, J. L., Stroer, B. S., Firszt, J. B., & Novak, M. A. (1991). Independent evaluation of the speech perception abilities of children with the Nucleus 22-channel cochlear implant system. *Ear & Hearing, 12* (Suppl.), S66-S80.

Osberger, M. J., Robbins, A. M., Miyamoto, R. T., Berry, S. W., Myres, W. A., Kessler, K. S., & Pope, M. L. (1991). Speech perception abilities of children with cochlear implants, tactile aids, or hearing aids. *American Journal of Otology, 12* (Suppl.), S105-S115.

Quittner, A. L., & Steck, J. T. (1991). Predictors of cochlear implant use in children. *American Journal of Otology, 12,* 89-94.

Robbins, A. M., Renshaw, J. J., Miyamoto, R. T., Osberger, M. J., & Pope, M. J. (1988). Minimal pairs test. Indianapolis, IN: Indiana University School of Medicine.

Sehgal, S. T., Kirk, K. I., Svirsky, M. A., & Miyamoto, R. T. (submitted). The effects of processor strategy on the speech perception performance of pediatric nucleus multichannel cochlear implant users. *Ear & Hearing.*

Skinner, M. W., Holden, L. K., Holden, T. A., Dowell, R. C., Seligman, P. M., Brimacombe, J. A., & Beiter, A. L. (1991). Performance of postlinguistically deaf adults with the Wearable Speech Processor (WSP III) and Mini Speech Processor (MSP) of the Nucleus multi-electrode cochlear implant. *Ear and Hearing, 12,* 3-22.

Skinner, M. W., Clark, G. M., Whitford, L. A., Seligman, P. M., Staller, S. J., Shipp, D. B., Shallop, J. K., Everingham, C., Menapace, C. M., Arndt, P. L., Antogenelli, T., Brimacombe, J. A., Pijl, S., Daniels, P., George, C. R., McDermott, H. J., & Beiter, A. L. (1994). Evaluation of a new spectral peak coding strategy for the Nucleus 22 channel cochlear implant system. *American Journal of Otology, 15,* 15-27.

Somers, M. N. (1991). Speech perception abilities in children with cochlear implants or hearing aids. *American Journal of Otology, 12* (Suppl.), S174-S178.

Staller, S. J., Beiter, A. L., Brimacombe, J. A., Mecklenburg, D. J., & Arndt, P. (1991). Pediatric performance with the Nucleus 22-channel cochlear implant system. *American Journal of Otology, 12* (Suppl.), S126-S136.

Staller, S. J., Dowell, R. C., Beiter, A. L., Brimacombe, J. A. (1991). Perceptual abilities of children with the Nucleus 22-channel cochlear implant. *Ear & Hearing, 12* (Suppl.), S34-S47.

Svirsky, M. A. (1996). Speech production and language development in pediatric cochlear implant users. Paper presented at the Annual Convention of the American Speech-Language-Hearing Association. Seattle, WA.

Waltzman, S., Cohen, N., Gomolin, R., Ozdamar, S., Shapiro, W., & Hoffman, R. (1995). Effects of short-term deafness in young children implanted with the Nucleus cochlear prosthesis. *Annals of Otology, Rhinology, & Laryngology, 104* (Suppl. 166), 341-342.

Waltzman, S. B., Cohen, N. L., Gomolin, R. H., Shapiro, W. A., Ozdamar, S. R., & Hoffman, R. A. (1994). Long-term results of early cochlear implantation in congenitally and prelingually deafened children. *American Journal of Otology, 12* (Suppl.), 9-13.

Waltzman, S. B., Cohen, N. L., & Shapiro, W. A. (1992). Use of a multichannel cochlear implant in the congenitally and prelingually deaf population. *Laryngoscope, 102,* 395-399.

Zwolen, T. A., Zimmerman-Phillips, S., Ashbaugh, C. J., Hieber, S. J., Kileny, P. R., & Telian, S. A. (1997). Cochlear implantation of children with minimal open-set speech recognition skills. *Ear and Hearing, 18,* 240-251.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 21 (1996-1997)
Indiana University

Predicting Open-Set Spoken Word Recognition Performance from Feature Identification Scores in Pediatric Cochlear Implant Users: A Preliminary Analysis[1]

Stefan Frisch and David B. Pisoni[2]

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405

---

[1] This work supported by NIH-NIDCD Training Grant DC00012 to Indiana University. We would like to thank Steve Chin, Paul Luce, and Ted Meyer for providing comments and criticism on this work.

[2] Also DeVault Otologic Research Laboratory, Department of Otolaryngology-Head and Neck Surgery, Indiana University School of Medicine, Indianapolis, IN 46202.

Predicting Open-Set Spoken Word Recognition Performance from Feature Identification Scores in Pediatric Cochlear Implant Users: A Preliminary Analysis

Abstract. This study is a first attempt at predicting open-set word recognition performance by pediatric cochlear implant users based on closed-set word identification scores. A software simulation of phoneme recognition was developed which uses feature identification scores from the Minimal Pairs Test (Robbins, Renshaw, Miyamoto, Osberger, and Pope 1988) to predict phoneme identification performance in open-set word recognition tasks. Simulated phoneme identification performance was then applied to word recognition on the Phonetically Balanced Kindergarten (PBK) test (Haskins 1949) and the Lexical Neighborhood Test (LNT) (Kirk, Pisoni, and Osberger 1995). Simulations were carried out using behavioral data from two samples. These samples were the best and worst performers on the PBK, studied in Pisoni, Svirsky, Kirk, and Miyamoto (this volume). The simulation generated good estimates of actual performance on the LNT for both populations when performance was scored by phonemes correct. However, when performance was scored by words correct, actual performance on the LNT was much better than predicted based on phoneme recognition. For the PBK, the simulation performed significantly better on phonemes correct for both populations. Differences between actual and predicted performance on words correct were small for children who performed well on the PBK, but children who performed poorly on the PBK scored much worse than predicted by the model. Like normal hearing adults, some exceptionally good pediatric cochlear implant users recognize words in the context of other words that they have in their lexicons. The phoneme confusion model of word recognition does not adequately predict word identification performance in open-set tests. In addition, differences between the two samples suggest that the best performers on the PBK have a more developed mental lexicon than the worst performers, and are familiar with more words on that test. Our simulations reveal the importance of using appropriate tests of word recognition for the clinical population and that closed-set and open-set word recognition performance are not transparently related. These simulations also highlight the important role of the mental lexicon in open-set word recognition.

Introduction

This study is a first attempt at predicting open-set word recognition performance on the Phonetically Balanced Kindergarten (PBK) test (Haskins 1949) and Lexical Neighborhood Test (LNT) (Kirk, Pisoni, and Osberger 1995) by pediatric cochlear implant users based on closed-set feature identification scores. Our study focuses on two groups of pediatric cochlear implant users, first examined in Pisoni, Svirsky, Kirk, and Miyamoto (this volume), who were the best and worst performers on the PBK test. These two groups are of interest for several reasons. First, they provide a wide range of performance levels over which to evaluate the relation between feature identification and open-set word recognition. Second, since these groups consist of prelingually deafened children, they have acquired spoken language through their implant, and their spoken word recognition performance provides an important test of the generality of current hypotheses about language processing in normal hearing populations. Third, Pisoni et al. (this volume) conclude that these two groups are qualitatively different in their language abilities and we wish to investigate whether this difference can be predicted from quantitative differences in their test performance with a single model, or whether different models are needed for each population. Finally, Pisoni et al (this volume) found no significant correlations between feature identification and open-set word recognition for these groups at one year post-implant, suggesting that a more detailed analysis is required to determine if there is a reliable relation between feature identification and open-set word recognition.

We develop a probabilistic model of phoneme recognition based on performance in feature identification in the Minimal Pairs Test (Robbins, Renshaw, Miyamoto, Osberger, and Pope 1988). Phoneme confusions were predicted by assessing the contrastiveness of different types of linguistic features in the phoneme inventory of English. A probabilistic confusion matrix was created for each participant based on their performance on the feature identification task. The confusion matrix was then applied as a simple model of word recognition, which predicts that a word is recognized only on the basis of independently identifying the individual phonemes in the word. This provided a useful benchmark to compare open-set word recognition performance across several conditions.

It is well known that normal hearing adults recognize words by accessing a mental lexicon, and that the frequency and density of words phonologically similar to the target word influences performance (Goldinger, Luce, and Pisoni 1989; Luce, Pisoni, and Goldinger 1990; Luce and Pisoni 1998). Evidence that normal hearing children and pediatric cochlear implant users have analogous lexical organization has also been found (Kirk et al. 1995; Logan 1992). Therefore, it is not expected that a model of spoken word recognition based solely on phoneme recognition would accurately predict open set recognition performance. However, the way in which the model fails to fit performance data provides additional evidence for three conclusions drawn elsewhere. First, pediatric cochlear implant users are sensitive to the phonetic similarity of words and employ this knowledge in the word recognition task (Kirk et al 1995). Second, there are substantial differences in the lexical characteristics of words on the PBK and LNT tests which suggest that the PBK is an extremely difficult test for pediatric cochlear implant users which significantly under-predicts their ability to recognize familiar spoken words in isolation (Kirk et al. 1995; Meyer and Pisoni this volume). Third, children who perform exceptionally well on the PBK test (the 'Stars') have made significant progress in developing a mental lexicon, a crucial step in acquiring spoken language (Pisoni, Svirsky, Kirk, and Miyamoto this volume). The phoneme confusion model is an important first step in developing a more complex, psycholinguistically valid, model of open-set word identification. The primary goal of this paper is to lay the groundwork for future research modeling open-set spoken word recognition by pediatric cochlear implant users based on current theories of language processing.

Predicting Phoneme Confusions from Linguistic Features

The Minimal Pairs Test is a two-alternative forced-choice word identification test containing words which contrast on a single phoneme (Robbins et al. 1988). The set of minimally contrastive phonemes covers a variety of linguistic features which represent the basic contrasts of English. In the Minimal Pairs Test, these contrasts are grouped into five broad phonetic categories. These categories and the phonemic contrasts used to represent each category are shown in Table 1.[3] Presumably, failure to differentiate contrasts for place of articulation for one consonant pair (e.g., p/k) indicates that discrimination for similar pairs (e.g., b/g, k/t) is also difficult. While some of the contrasts which have been grouped into a single category may have very different acoustic realizations (e.g., place in p/k versus ʃ/f, manner in m/b versus f/p, voicing in p/b versus v/f), this study assumes for simplicity that all contrasts within a class are equivalent and that performance on any one contrast can be predicted from the average performance over the phoneme pairs that test that contrast.

---

[3] Note that some of the contrasts in the Minimal Pairs Test involve more than a single feature category. The t/ʃ contrast is both a manner and (minor) place difference. The u/I contrast is both a vowel place and tense/lax difference. The i/ɔ contrast is both a vowel place and vowel height difference. We assume the confounded features do not greatly affect the estimates of feature identification performance.

Table 1

Phonemic contrasts in the Minimal Pairs Test.

Consonant Contrasts			Vowel Contrasts	
Place	Manner	Voicing	Vowel Place	Vowel Height
p/k	ʃ/t	p/b	ɪ/ə	i/ɔ
p/t	m/b	k/g	u/i	u/o^{u}
ʃ/f	f/p	v/f	u/ɪ	æ/i

In order to make predictions of phoneme confusions from the Minimal Pairs Test, it is necessary to decide explicitly whether a failure to identify some contrast will result in confusions between a particular pair of phonemes. For example, if place of articulation for a particular pair of consonants cannot be discriminated, are the remaining manner and voicing features sufficient to differentiate the consonants? In linguistic theory, this is referred to as 'distinctness' (Stanley 1967). In order to answer this question, we turn to a set-theoretic model of the phonemic inventory originally developed in Broe (1993), which makes explicit the contrastiveness of phonemic representations (see also Frisch 1996).

Representing Phonological Contrast: Structured Specification

The theory of structured specification represents the phoneme inventory of a language using the hierarchy of the set of 'natural classes' defined by the distinctive features of that inventory. A 'natural class' is simply a set of phonemes defined by a conjunction of features. Structured specification makes the natural classes an explicit part of the representation of the inventory, and in so doing makes explicit the patterns of contrast and distinctiveness given by the features. A hierarchy among natural classes arises because the larger, more general natural classes contain the smaller, more specific ones. This hierarchy will be illustrated with a simple example: A five vowel inventory {a, e, i, o, u}. A typical set of features for this inventory, grouped into categories, is given in (1).[4]

---

[4] Note that all features used are monovalent, and thus are either present (+) or not. Structured specification makes the notational distinction between the more familiar bivalent (+/-) and monovalent or multivalent features irrelevant. Monovalent features are used here for consistency with Frisch (1996).

(1)

		a	e	i	o	u
VOWEL HEIGHT	[high]			+		+
	[mid]		+		+	
	[low]	+				
VOWEL PLACE	[front]		+	+		
	[back]	+			+	+

From these feature assignments, we can generate a variety of natural classes, partially ordered by set containment. Examples are shown in (2). Note from (2a) that the natural classes provide information about the contrastiveness of features. The set of [+low] segments, {a}, is entirely contained in the set of [+back] segments, {a, o, u}. This means that [+low] ⇒ [+back], or that vowel place is not contrastive for low vowels. In other words, there is no natural class {[+low], [+back]} distinct from {[+low]}. Examining the natural classes reveals which features are required to make phonemes distinct, and which features are redundant.

(2) a. {[+back]} = {a, o, u} ⊃ {[+low]} = {a}
b. {[+front]} = {i, e} ⊃ {[+mid], [+front]} = {e}
c. {[+high]} = {i, u} ⊃ {[+high], [+front]} = {i}

In structured specification theory, the phoneme inventory is represented as a lattice of all of the natural classes for that inventory given the set of feature assignments (Broe 1993). The lattice of the five vowel inventory is shown in Figure 1. The top node in the lattice is the entire inventory. The bottom node is the empty set. Each node in between represents a natural class. Each natural class is given with both the phonemes in that class, and the features which define that class. Lines from larger natural classes to smaller natural classes indicate set containment. All of the containment relations in (2) are represented structurally in the lattice. The row of nodes just above the bottom are natural classes containing each individual segment. The fact that each segment occupies its own natural class indicates that the distinctive features used to describe this inventory are sufficient to individuate the phonemes: Each phoneme has a distinct feature set which identifies it.

Insert Figure 1 about here

We can use the natural classes and the lattice representation to predict phoneme confusions when contrasts are not discriminable. For example, suppose that no VOWEL HEIGHT features are discriminable. Figure 2 shows the five vowel inventory with none of the VOWEL HEIGHT features. The phonemes are no longer individuated, and there are two minimal natural classes: {e, i} and {a, o, u}. With no VOWEL HEIGHT features, /e/ and /i/ are confusable, and /a/, /o/, and /u/ are confusable.

Insert Figure 2 about here

Figure 3 shows the five vowel inventory with the VOWEL PLACE features removed. In this case, /a/ is still individuated, as it is the only [+low] vowel in the inventory. The fact that /a/ is [+back] is redundant, so the loss of that information does not affect its distinctness. The phonemes /e/ and /o/, and /i/ and /u/, are confusable if there are no VOWEL PLACE contrasts.

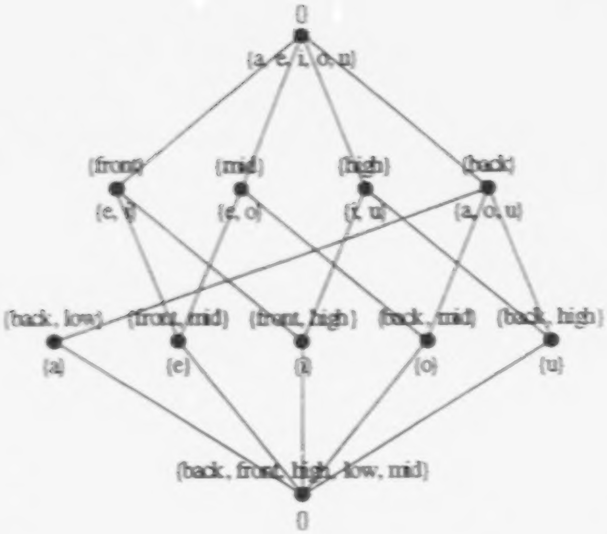

{}
{a, e, i, o, u}
{front}
{mid}
{high}
{back}
{e, i}
{e, o}
{i, u}
{a, o, u}
{back, low}
{front, mid}
{front, high}
{back, mid}
{back, high}
{a}
{e}
{i}
{o}
{u}
{back, front, high, low, mid}
{}

Figure 1. Lattice of natural classes of the five vowel inventory {i, e, a, o, u}.

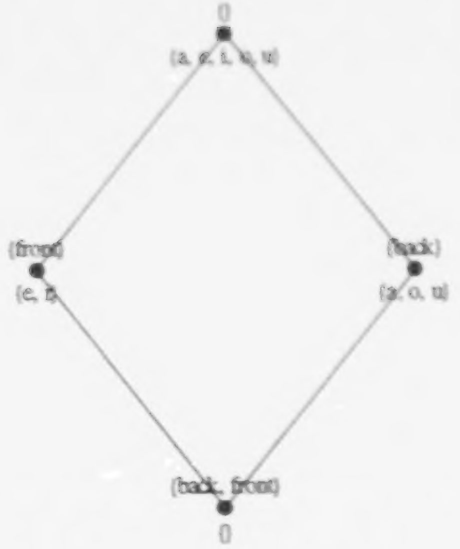

{}
{a, e, i, o, u}
{front}
{e, i}
{back}
{a, o, u}
{back, front}
{}

Figure 2. Lattice of natural classes of the five vowel inventory {i, e, a, o, u} with no Vowel Height features.

Insert Figure 3 about here

Generating Categorical Confusion Matrices

Once the confusable phonemes are known, the pattern of confusability can be represented using a confusion matrix. For simplicity, it is assumed that when several phonemes are confusable, each is equally likely to be identified as the perceived phoneme.[5] So the probability of any particular confusion is inversely related to the size of the set of possible confusions. The confusion matrix provides a convenient display for the probability of confusions in an entire inventory when a particular contrast is lost. Tables 2 and 3 show confusion matrices for the five vowel inventory when VOWEL HEIGHT or VOWEL PLACE features are removed. The intended phoneme is given in the left column, the perceived phoneme is given across the top. Each number in the table represents the probability that the intended phoneme is heard as the perceived phoneme. A phoneme is always assumed to be confusable with itself. In other words, there is a possibility that the correct phoneme will be identified even if a contrast is not detectable.

Table 2

Confusions in the five vowel inventory when VOWEL HEIGHT features are removed

	Perceived				
Intended	i	e	a	o	u
i	0.5	0.5	0	0	0
e	0.5	0.5	0	0	0
a	0	0	0.33	0.33	0.33
o	0	0	0.33	0.33	0.33
u	0	0	0.33	0.33	0.33

Table 3

Confusions in the five vowel inventory when VOWEL PLACE features are removed

	Perceived				
Intended	i	e	a	o	u
i	0.5	0	0	0	0.5
e	0	0.5	0	0.5	0
a	0	0	1.0	0	0
o	0	0.5	0	0.5	0
u	0.5	0	0	0	0.5

---

[5] A more realistic model would weight confusions by the frequency of the phonemes involved. Including a phoneme frequency effect requires a number of additional assumptions about whether to use lexical frequency or frequency in running speech, and whether frequency should be position independent or sensitive to position in word and position in syllable. These factors deserve independent consideration and are left as open research questions.

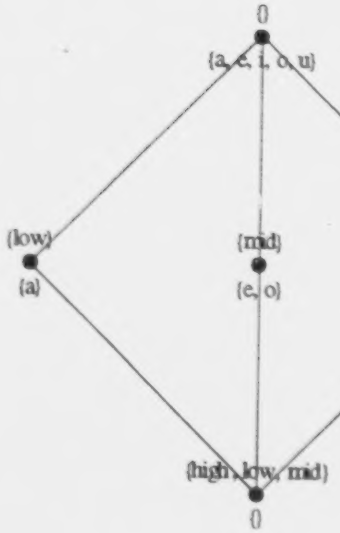

{}
{a, e, i, o, u}
{low}
{a}
{mid}
{e, o}
{high}
{i, u}
{high, low, mid}
{}

Figure 3. Lattice of natural classes of the five vowel inventory {i, e, a, o, u} with no Vowel Place features.

Confusions for more complicated phoneme inventories, like the full set of vowels and consonants in English, can be predicted using the lattice representation in exactly the same way. For larger inventories with many more phonemes and features, the natural classes are particularly useful. Unlike the five vowel inventory, confusions for the full inventory are much more difficult to determine by simple inspection of the feature matrix. The phoneme inventory and feature sets used for the computational model in this paper are given in the Appendix. The consonant features are divided into five categories of contrast, ARTICULATOR features, PLACE features, STRICTURE features, MANNER features, and LARYNGEAL features. The ARTICULATOR and PLACE features fall under the contrasts represented by the Place category in the Minimal Pairs Test. The STRICTURE features and MANNER features are represented by the Manner category in the Minimal Pairs Test. The LARYNGEAL features are represented by the Voicing category. The vowel features are divided into three categories of contrast, VOWEL PLACE, VOWEL HEIGHT, and VOWEL MANNER. The VOWEL PLACE and VOWEL HEIGHT features correspond to the same categories in the Minimal Pairs Test. There is no correspondent to the VOWEL MANNER features in the Minimal Pairs Test. For the simulations in this paper, VOWEL MANNER features are assumed to be perfectly discriminated, except for the [+static] and [+dynamic] features that individuate diphthongs from monophthongs. Since the differences between the parts of a diphthong correspond to dynamic changes in both place and height features, when neither place or height is discriminable, the contrast between diphthongs and monophthongs is assumed to be lost.

Predicting Phoneme Identification

Distinctive features and natural classes provide several ways of understanding the loss of contrast when a feature is categorically absent from the representation. From a perceptual perspective, this corresponds to a lack of ability to identify the perceptual cues which signal the contrast. However, behavioral measures rarely reveal a complete absence of distinctness. Instead, identification is imperfect or unreliable for particular contrasts. For our simulation, imperfect identification is treated as a probabilistic determinant of whether or not a particular featural contrast is perceived. Performance on the Minimal Pairs Test can thus be used to define a probability distribution of confusions for each phoneme. The method for generating this distribution is discussed in this section.

The Minimal Pairs Test is a two-alternative forced-choice word identification task (Robbins et al. 1988). The child is presented with one word on each trial and responds by selecting one of two alternative pictures. The first step in converting performance on such a task to a prediction of open-set word recognition performance is to transform the two-alternative word identification score into an estimate of open-set feature identification performance. There are no studies we know of which have evaluated the relation between closed-set word recognition and open-set feature identification in this way. There are some related results comparing open set and closed set word recognition. Black (1957) found that there was no interaction between open versus closed test format and changes in presentation signal level on intelligibility of spoken words. From these findings, he argued that the only difference in performance is an adjustment in base rate and therefore that closed-set tasks should be used clinically due to their ease of implementation and scoring. More recently, Sommers, Kirk, and Pisoni (1997) found differences between open-set and closed-set word recognition in effects of talker variability and lexical neighborhood effects. No effects of talker or neighborhood were found for closed set word recognition tasks. The data modeled in this paper do not include talker manipulations, so this factor is irrelevant for the current analysis. The phoneme-based model of word recognition developed in this paper does not include a lexical component, so differences in lexical effects cannot be captured (see below). However, it is important to point out here that the assumption of equivalence between open-set and closed-set speech intelligibility tests as suggested by Black may be unwarranted, and in fact misleading if we are to understand the perceptual processes used to recognize words in these two different contexts.

In the present analysis, open-set feature identification performance is assumed to be equivalent to closed-set word identification performance for the words on the Minimal Pairs Test, with the advantage of the restricted choices in a closed-set task factored out. The equation for estimated open-set feature identification performance is given in (3). This equation merely scales the performance from the closed set range of chance performance to 100% to the open set range of 0% to 100%. For example, two-alternative closed-set performance of 50% corresponds to 0% open-set performance (chance); closed-set performance of 75% corresponds to 50% open-set performance.

(3) open set = (closed set - chance) ÷ (100% - chance)

The Minimal Pairs Test provides identification scores for individual feature contrasts. The simplest assumption for identification of several contrasts is that the identification of each contrast is independent. Under this assumption, the probability of simultaneously identifying multiple contrasts is the product of the individual contrast probabilities. For consonants, there are three contrasts and for vowels there are two. Assuming the probability of identifying PLACE, MANNER, and VOICING are p, m, and v, respectively, and that the probability of identifying VOWEL PLACE and VOWEL HEIGHT are vp and vh, the distribution of independent possibilities for feature detection in vowels and consonants is given in Table 4. Note that the representations of structured specification used above for predicting confusions when a single contrast is lost apply equally well when multiple contrasts are lost. A lattice can be generated with any set of features removed, and the resulting minimal natural classes represent the phonemes which are no longer distinct when those features are lost.

Table 4

Probability distribution of identification performance based on categories in the Minimal Pairs Test

Consonants		Vowels	
Contrasts	Probability	Contrasts	Probability
None	(1-p)·(1-m)·(1-v)	None	(1-vp)·(1-vh)
Place	p·(1-m)·(1-v)	Vowel Place	vp·(1-vh)
Manner	(1-p)·m·(1-v)	Vowel Height	(1-vp)·vh
Voicing	(1-p)·(1-m)·v	All	vp·vh
Place & Manner	p·m·(1-v)		
Place & Voicing	p·(1-m)·v		
Manner & Voicing	(1-p)·m·v		
All	p·m·v		

For example, with performance of 60% on Place, 70% on Manner, and 80% on Voicing on the Minimal Pairs Test, the predicted probability of identifying Place and Manner but not Voicing would be

$p \cdot m \cdot (1-v) = 0.2 \cdot 0.4 \cdot (1-0.6) = 0.032$. The probability of identifying no features would be $(1-0.2) \cdot (1-0.4) \cdot (1-0.6) = 0.192$.

Using the estimated probabilities of different outcomes shown in Table 4, we can generate a probabilistic distribution of contrast confusions for each phoneme. The probabilities for each possible outcome can then be used as weights for combining categorical confusion matrices of the sort given in Tables 2 and 3. The result is a probabilistic confusion matrix which takes into account the statistical reliability of feature identification for individual features and their combinations. This distribution is used below to predict phoneme confusions in open set word recognition. A sample probabilistic confusion matrix using the five vowel inventory will be worked out in detail to illustrate the computational procedure.

Suppose that two-alternative closed-set identification in a five vowel inventory is 60% for Vowel Place, and 80% for Vowel Height. Then open set identification is estimated to be 20% for Vowel Place, and 60% for Vowel Height. Based on these values, the estimated probability distribution for detecting contrasts is given in (4).

(4)

Contrasts maintained	Estimated probability
None	0.32
Vowel Place	0.08
Vowel Height	0.48
All	0.12

Tables 2 and 3 provide confusion matrices for the cases where only Vowel Place and only Vowel Height are contrastive. The confusion matrices when neither or both are contrastive are obvious. If neither is contrastive, then all phonemes are confusable, as there are no features remaining. For a five vowel inventory, the probability of a particular perceptual result at random is 0.2. When both Vowel Place and Vowel Height are contrastive, then no confusions result. Each probability in the distribution in (4) is multiplied by the appropriate confusion matrix, as shown in (5). Equation (5) is presented in the matrix notation of linear algebra. To determine the resulting matrix, each cell in each matrix is multiplied by the appropriate probability (weight), and corresponding cells are then summed across matrices. For example, the upper left cell, representing confusions between /i/ and itself, has overall probability of $0.32 \cdot 0.2 + 0.48 \cdot 0.5 + 0.08 \cdot 0.5 + 0.12 \cdot 1 = 0.464$. The resulting probabilistic confusion matrix is given in Table 5. This matrix reflects the initial probabilities used in the example. The example listener was better at detecting height differences than place differences, and the matrix predicts more place confusions than height confusions.

(5)

$$0.32 \cdot \begin{vmatrix} 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \\ 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \\ 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \\ 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \\ 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \end{vmatrix} + 0.48 \cdot \begin{vmatrix} 0.5 & 0 & 0 & 0 & 0.5 \\ 0 & 0.5 & 0 & 0.5 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0.5 & 0 & 0.5 & 0 \\ 0.5 & 0 & 0 & 0 & 0.5 \end{vmatrix} + 0.08 \cdot \begin{vmatrix} 0.5 & 0.5 & 0 & 0 & 0 \\ 0.5 & 0.5 & 0 & 0 & 0 \\ 0 & 0 & 0.33 & 0.33 & 0.33 \\ 0 & 0 & 0.33 & 0.33 & 0.33 \\ 0 & 0 & 0.33 & 0.33 & 0.33 \end{vmatrix} + 0.12 \cdot \begin{vmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{vmatrix}$$

Table 5

Example confusion matrix for the five vowel inventory

	Perceived				
Intended	i	e	a	o	u
i	0.464	0.104	0.064	0.064	0.304
e	0.104	0.464	0.064	0.304	0.064
a	0.064	0.064	0.691	0.091	0.091
o	0.064	0.304	0.091	0.451	0.091
u	0.304	0.064	0.091	0.091	0.451

Confusion matrices for the inventories of consonants and vowels in English can be generated following the exact same algorithm. In the case of consonants, eight probability weighted confusion matrices are added together based on the estimated probabilities of feature detection in the left column of Table 4. For vowels, four such matrices are added (as in the example, but with the full vowel inventory of English) based on the estimated probabilities of feature detection in the right column of Table 4.

A Phoneme Confusion Model of Spoken Word Recognition

Given a model of phoneme confusion for open-set tasks, the simplest model of spoken word recognition applies the phoneme confusion model on each phoneme in the word and treats the result as the recognized word. This model does not employ any lexical knowledge, as the perceptual result is not matched to an internalized representation of words in the mental lexicon. This model is certainly unrealistic as a model of normal spoken word recognition. Boothroyd and Nittrouer (1988) concluded that such a model was appropriate for non-word recognition of normal hearing young adult listeners. They found that recognition of CVC non-word syllables was related to phoneme recognition by a simple formula: $p_w = p_p^3$, where p^w is the probability of correct word recognition and p_p is the probability of correct phoneme recognition. For words, they found that $p_w = p_p^j$ where j is significantly less than 3 (approximately 2.5), which implies that word recognition is better than predicted by phoneme recognition. Nittrouer and Boothroyd (1990) replicated this finding with normal hearing older adult listeners (mean age 72) and with normal hearing 5 year olds. Even though the overall performance of children and older adults was lower in percent correct, the j factor representing the contribution of lexical knowledge was not significantly different from the young adults. Rabinowitz, Eddington, Delhorne, and Cuneo (1992) observed similar results with adult cochlear implant users. They found $j = 2.65$, again well below 3. Note that the Rabinowitz et al. j coefficient, slightly higher than that found by Boothroyd and Nittrouer, is not directly comparable as different tests were used to determine word and phoneme recognition scores. Together, these studies show that populations which acquired language normally take advantage of the structure of the mental lexicon when identifying words.

It has also been found that open set word recognition is influenced by competition of phonemically similar candidates in the mental lexicon, in both normal hearing adults and pediatric cochlear implant users (Luce and Pisoni 1998; Kirk et al. 1995). This finding has been replicated several times for normal hearing adults (see Luce, Pisoni, and Goldinger 1990) but has only recently been replicated in the pediatric cochlear implant population. In particular, Pisoni et al. (this volume) have claimed that there may be differences among members of this population which is relevant to modeling their open-set word recognition performance. They claim that the best performers on the PBK, the so-called 'Stars,' have internalized a mental lexicon of words which is crucial for language development. The worst performers on the PBK, called 'Controls' in Pisoni et al. (this volume), may not have developed such an internalized store. Thus, it may be the case that the phoneme confusion model of word recognition is appropriate for the 'Control' population, but not the 'Stars' or normal-hearing adults.

A sample application of the phoneme confusion model will now be presented, to give the reader a better feel for the predictions and results. The example is based on the average performance on the Minimal Pairs Test for the Stars group at two years post-implant. Mean two-alternative feature identification scores for this group are: Place 71.1%, Manner 80.1%, Voicing 70.3%, Vowel Place 92.7%, Vowel Height 94.2%. Fifty iterations of open set phoneme confusions for the word *please* /pliz/ were generated based on these identification scores. The first fifteen simulated words are given in (6). The simulation perceived the entire sequence of phonemes /pliz/ correctly in 6 out of 50 iterations (12% correct). When scored by phonemes correct, 125 out of 200 phonemes were perceived correctly (62.5%). Qualitatively, the simulation performed much as expected given the Minimal Pairs Test identification scores. The vowel was rarely confused (42 correct out of 50, 84%). For the initial stop /p/, Manner was most frequently preserved (34 out of 50, 68%) followed by Voicing (32 out of 50, 62%) and Place (30 out of 50, 60%).

(6) fniz
plið
rliz
dmiz
bliʒ
pviʃ
mlen
pliʒ
pniz
pdis
pzoᶦz
wliz
ʃliz
tlez
ʃlaz

Simulation and Results

The procedure demonstrated here for a single word using average performance data on the Minimal Pairs Test was applied to the word lists in the PBK (Haskins 1949) and LNT (Kirk et al. 1995) tests based on individual performance data from the two groups of subjects in Pisoni et al. (this volume), the 'Stars' and 'Controls.'

Participants

The data used in the computational simulation were originally obtained from a longitudinal study of 160 deaf children. Two groups, the 'Stars' and 'Controls,' were selected based on their performance on the PBK test. The 'Stars' were children who scored in the upper 20% on the PBK test. The 'Controls' scored in the bottom 20% on the PBK. Subjects in both groups were all prelingually deafened, but the groups differed on a variety of demographic variables. A more complete description of this population can be found in Pisoni et al. (this volume). In this simulation, data were used from one year, two years, and three years post-implant for each group. These data provide a range of performance on the Minimal Pairs Test, and also allow assessment of longitudinal changes in actual performance in comparison to the model. Since the Minimal Pairs Test was used to predict performance on the PBK and LNT, only participants with Minimal Pairs Test scores were used. In addition, while all the participants selected had PBK scores as part of the selection criteria, not all participants had LNT scores. The actual number of participants simulated in each group are given in Table 6.

Mean performance for each group on the Minimal Pairs Test is given in Figures 4 and 5. Figure 4 shows performance for consonant contrasts, and Figure 5 shows performance for vowel contrasts. Notice that, among the data for the three longitudinal samples for each group, there is a wide range of levels of performance to be examined.

Table 6

Number of participants simulated in the experiment

	Group	PBK	LNT Easy	LNT Hard
Stars	TY1	20	14	12
	TY2	24	17	16
	TY3	20	15	15
Controls	BY1	13	1	0
	BY2	15	7	5
	BY3	15	8	2

Insert Figure 4 about here

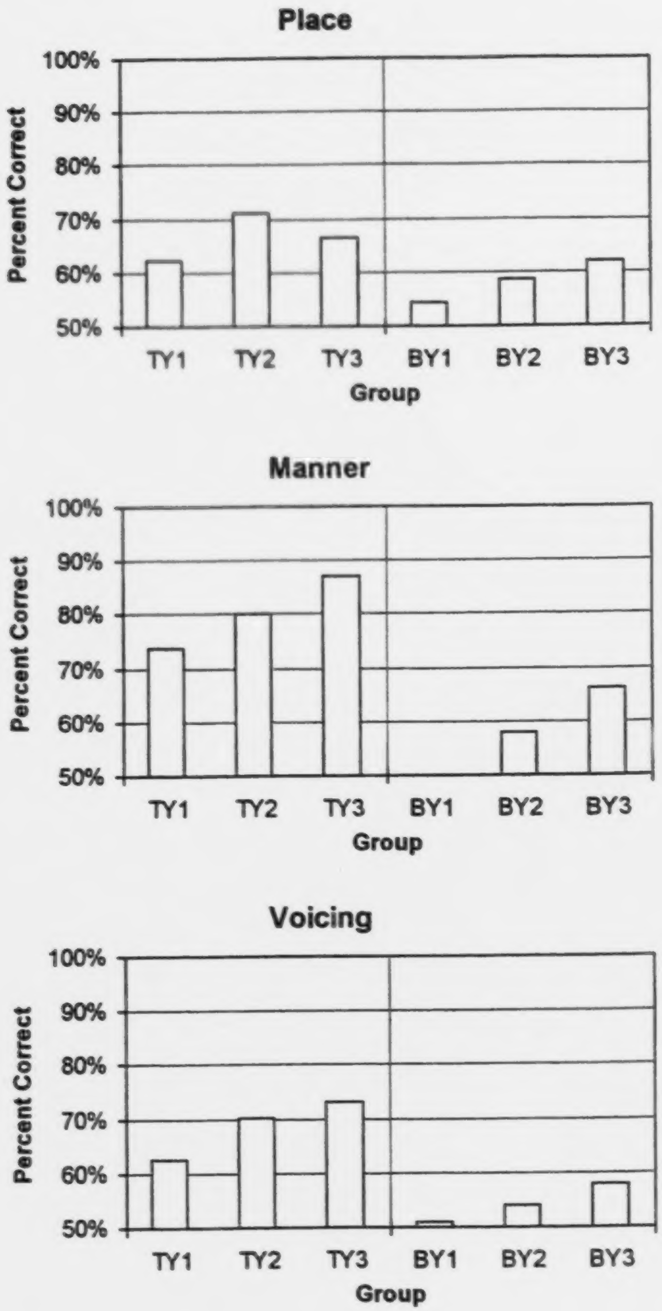
Insert Figure 5 about here

Material

The simulation performed phoneme by phoneme recognition for the 150 words on the PBK and for the 50 'easy' and 50 'hard' words on the LNT. On the LNT test, easy and hard words differ in their lexical characteristics (Kirk et al. 1995). Hard words have more high frequency competitors than easy words and are more difficult to recognize due to greater lexical competition from phonetically similar words. These lexical factors are irrelevant for the phoneme confusion model of word recognition, as is shown below.

Methods

Performance on the Minimal Pairs Test was used to generate a confusion matrix for vowels and consonants of the type shown in Table 5 for each participant. The confusion matrix was then applied phoneme by phoneme as demonstrated above to 50 repetitions of each word from the PBK, LNT easy, and LNT hard word lists. The simulated perceived words were scored for percent correct words and phonemes. Since many of the simulations were run for very few participants, statistically significant differences between simulation performance and actual performance was tested across participants in each category using sign tests.


Place

Manner

Voicing

Figure 4. Mean performance on the Minimal Pairs Test for consonant place, manner, and voicing. The 'Stars' are shown on the left, the 'Controls' are shown on the right.

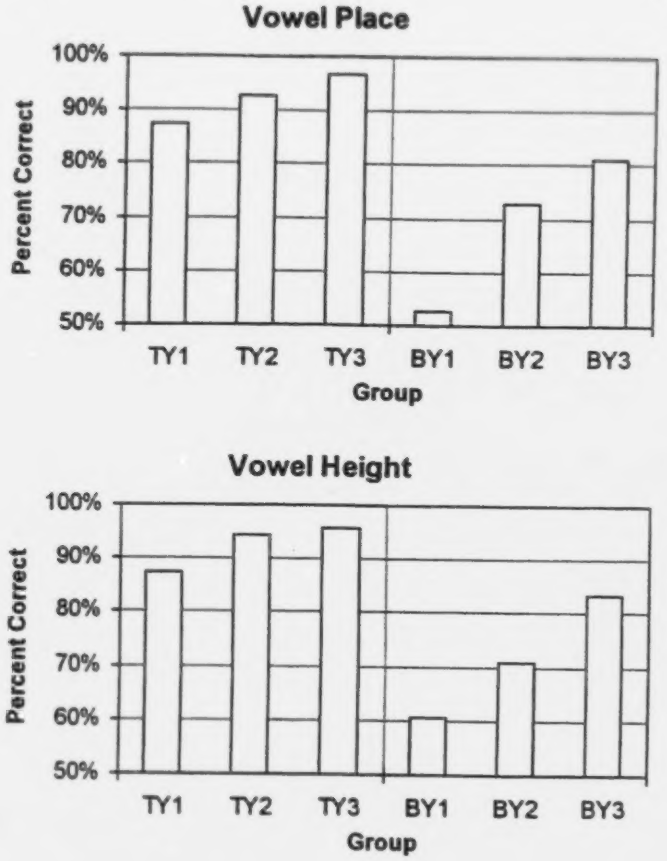


Figure 5. Mean performance on the Minimal Pairs Test for vowel place and vowel height. The 'Stars' are shown on the left, the 'Controls' are shown on the right.

Results

Mean actual and predicted performance for each group on the PBK test, scored in percent correct phonemes and words, is given in Figure 6. Note first that the overall developmental trends found in the actual performance data are reflected in the model by the use of the Minimal Pairs Test data. There are differences between the groups and longitudinally within groups in the participants' actual performance on the PBK which is also correctly reflected in their simulated performance. However, the phoneme confusion model of word recognition over predicts the number of correct phonemes identified for all participants. Also, in word recognition, the model over predicts the abilities of the 'Controls.' For the 'Stars,' however, the model provides a reasonably accurate prediction of word recognition up to two years post-implant. For the TY3 group, actual performance exceeds predicted performance. Interestingly, this is also the only group for which the model's performance was not significantly better than actual performance when scored with phonemes correct.

Insert Figure 6 about here

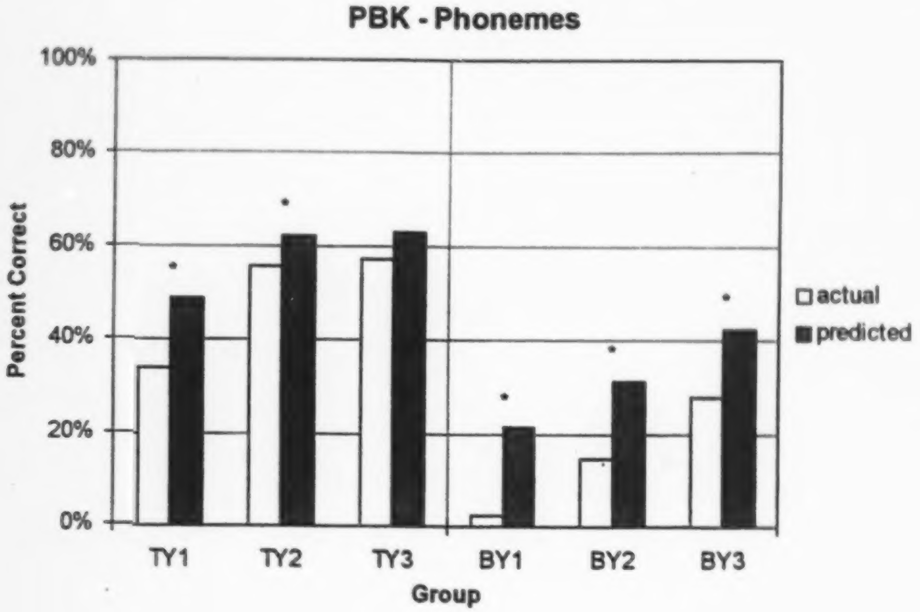
Mean actual and predicted performance for each group on the LNT easy word list, scored in percent correct phonemes and words, is given in Figure 7. Considering first the performance scored on phonemes correct, there are no significant differences between the model's performance and actual performance. When the scores of words correct are examined, we see that the model severely *under* predicts actual performance for all groups except BY1, which is data from a single subject who was unable to recognize any words in an open-set test format.

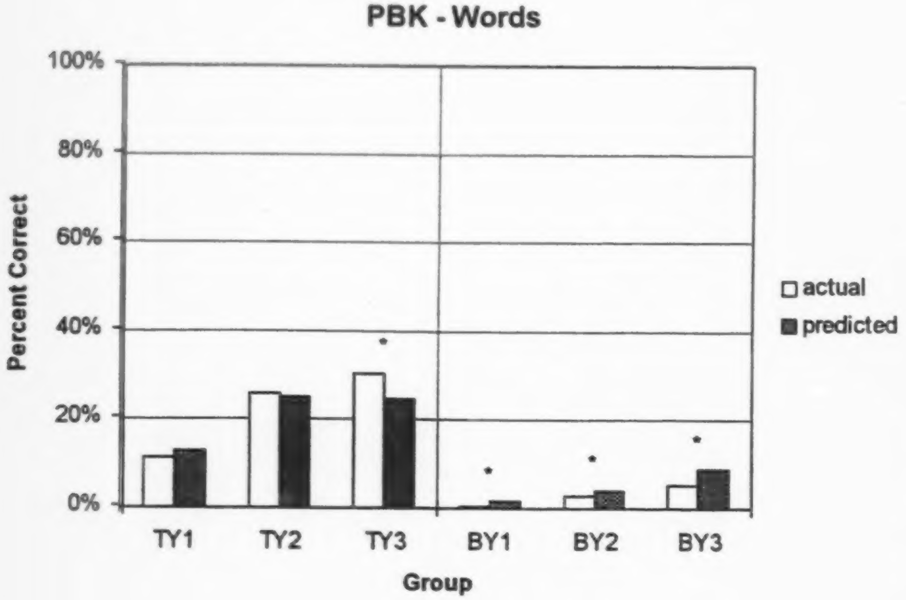
Comparing Figures 6 and 7, note that the predicted performance of the model on LNT easy words is not very different from its predicted performance on the PBK, scored either by words or phonemes. This indicates that there is nothing about the phonemic composition of these two word lists that could explain the large differences in observed performance for all subjects.

Insert Figure 7 about here

Mean actual and predicted performance for each group on the LNT hard word list, scored in percent correct phonemes and words, is given in Figure 8. Note first that there is no data for the BY1 group on this test. The pattern of results for the LNT hard words is much the same as for the LNT easy list, though the differences between actual and predicted performance on words is somewhat smaller for the 'Stars' group. This observed difference in performance is expected, since the hard word list was constructed of lexical items from denser neighborhoods of the lexicon. The phoneme confusion model of word recognition predicts roughly the same performance for the easy words as the hard words on this test, because there is no lexical lookup or competition in this model. There is very little data for the 'Controls' on this test, but the pattern is consistent with the pattern found for the 'Stars.'

Insert Figure 8 about here

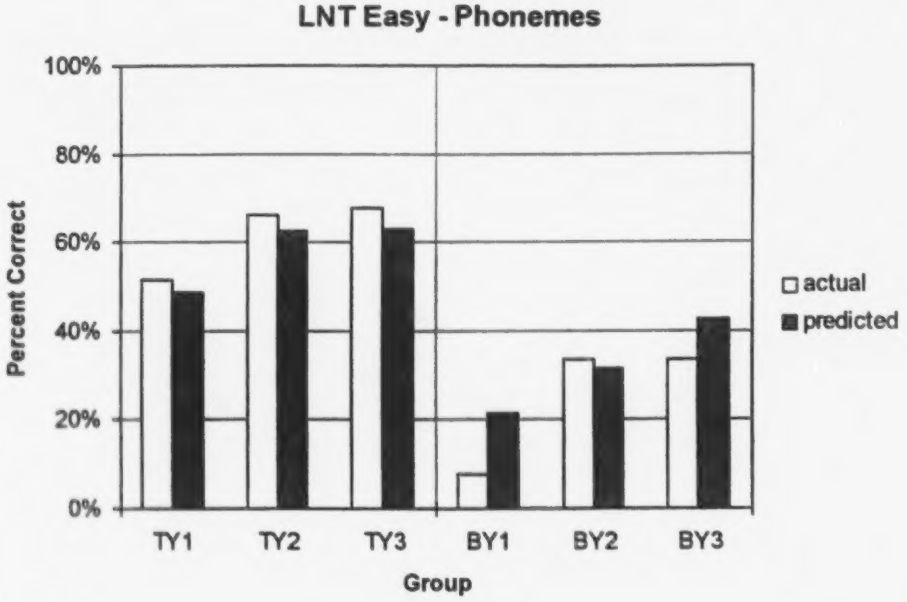


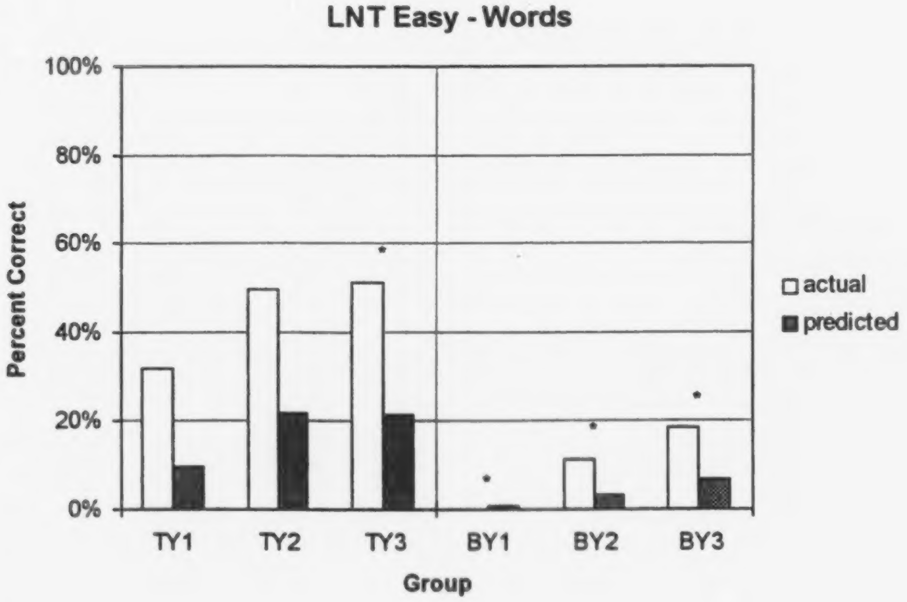


* $p < 0.05$

Figure 6. Comparison of phoneme and word performance on the PBK. The 'Stars' are shown on the left, the 'Controls' are shown on the right.

310

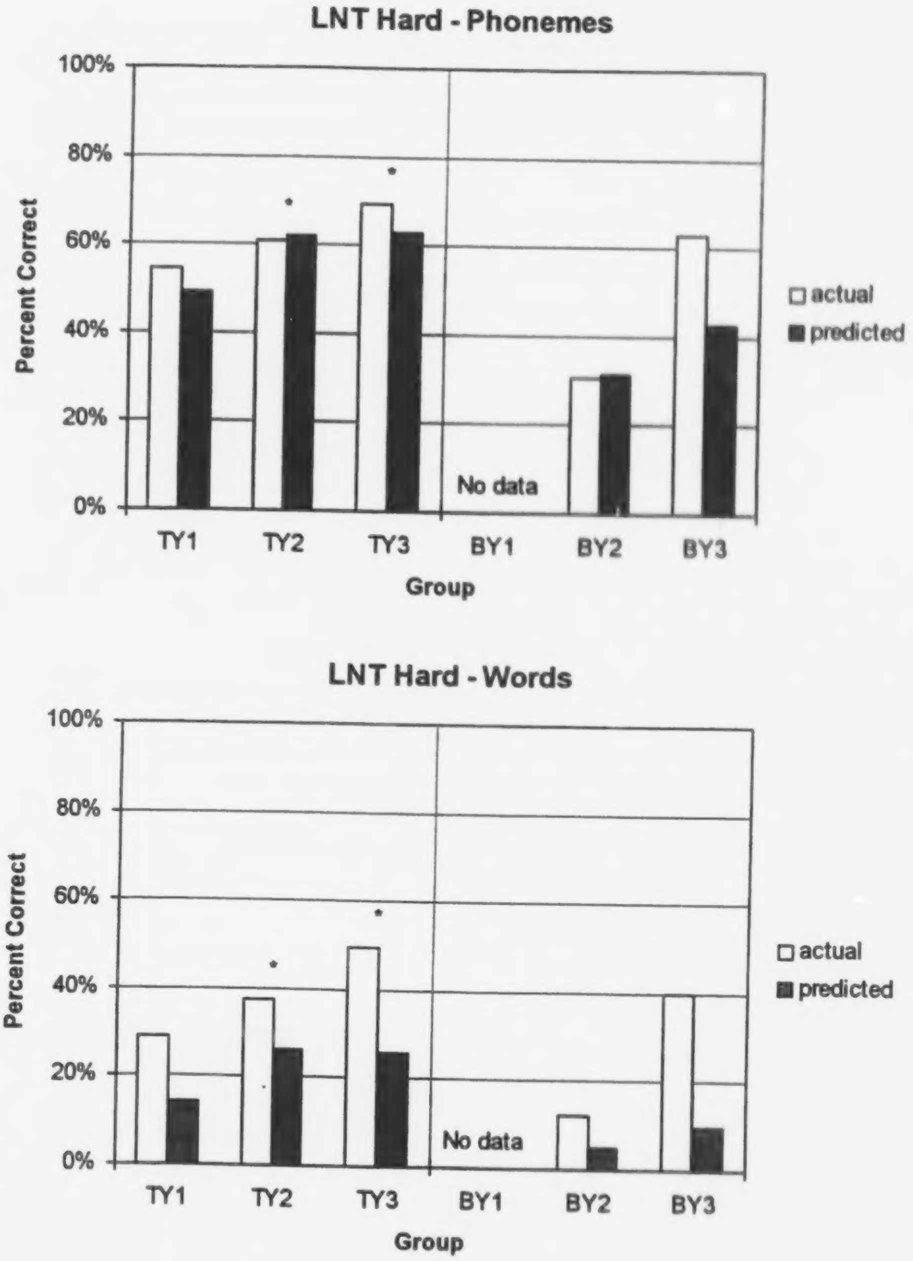




* $p < 0.05$

Figure 7. Comparison of phoneme and word performance on the LNT easy word list. The 'Stars' are shown on the left, the 'Controls' are shown on the right.

311



* $p < 0.05$

Figure 8. Comparison of phoneme and word performance on the LNT hard word list. The 'Stars' are shown on the left, the 'Controls' are shown on the right.

Discussion

Clearly, both groups of children performed much better than predicted by the phoneme confusion model in recognizing words on the LNT test. This suggests that, for words on the LNT word lists, pediatric cochlear implant users are comparing the degraded information they receive from their cochlear implant to internalized representations of lexical items in memory. Unlike the phoneme confusion model of word recognition, actual listeners can accommodate 'near misses', which are modeled here as perceptual mistakes in the recognition of a phoneme as in [bliz] or [pliʃ], and retrieve the appropriate item, /pliz/. However, when the LNT tests are scored by phonemes correct, the model does a reasonable job of predicting actual performance. This fact leads to two important conclusions. First, open-set phoneme identification can be approximately predicted by closed-set feature identification along the lines described in this paper. Second, since pediatric cochlear implant users perform much *better* than expected on recognizing words, which we would expect would increase their score on phonemes correct, it must be that on average partial information which points to an incorrect word candidate cancels out the gain in phoneme recognition given by a correct word candidate. For example, the near miss [bliz] might be matched in the lexicon as /pliz/ or /blid/. Choosing the correct word candidate increases the number of phonemes correct by one, but choosing the equally valid but incorrect word candidate results in two incorrect phonemes.

The present analysis also reveals a striking difference between the PBK and LNT word lists. On the PBK, observed accuracy in word recognition was no better than the model's performance for the 'Stars,' except at three years post implant. More interestingly, however, was the finding that observed accuracy scored in phonemes correct was much worse than predicted on the PBK. We might consider the possibility that the difference between the PBK and LNT lists is due to lexical factors. The LNT hard list is more difficult than the LNT easy list due to higher lexical frequency, lower neighborhood density, and lower lexical neighborhood frequency for the easy words (see Kirk et al. 1995). A similar difference was found for a fourth PBK list which Haskins (1949) found to be 'more audible' than the three lists which are regularly used (see Meyer and Pisoni this volume). To test this hypothesis, we carried out a computational analysis of all the words used on both the PBK and LNT lists. The comparison of lexical characteristics between the PBK words and LNT easy and hard words is shown in Figure 9. Mean lexical frequency, neighborhood density, and lexical neighborhood frequency for words in each test, computed using an on-line version of the Webster's pocket dictionary (see Nusbaum, Pisoni, and Davis 1984), is given. All three lexical factors indicate that the PBK words should be *harder* than LNT easy words, but easier than LNT hard words.

Insert Figure 9 about here

Kirk et al. provide a different account of the difference between the PBK and the LNT. They propose that even the most successful pediatric cochlear implant users with only a year or two of linguistic development are largely unfamiliar with the words on the PBK. In fact, Kirk et al. (1995) found that only 31% of the words on the PBK were present in a lexicon of child language derived from the CHILDES database by Logan (1992). If pediatric cochlear implant users are matching partial acoustic-phonetic information on words from the PBK which are not in their lexicons, then both their word and phoneme accuracy scores should be lower due to incorrect matches. This is exactly what we observed in our analysis.

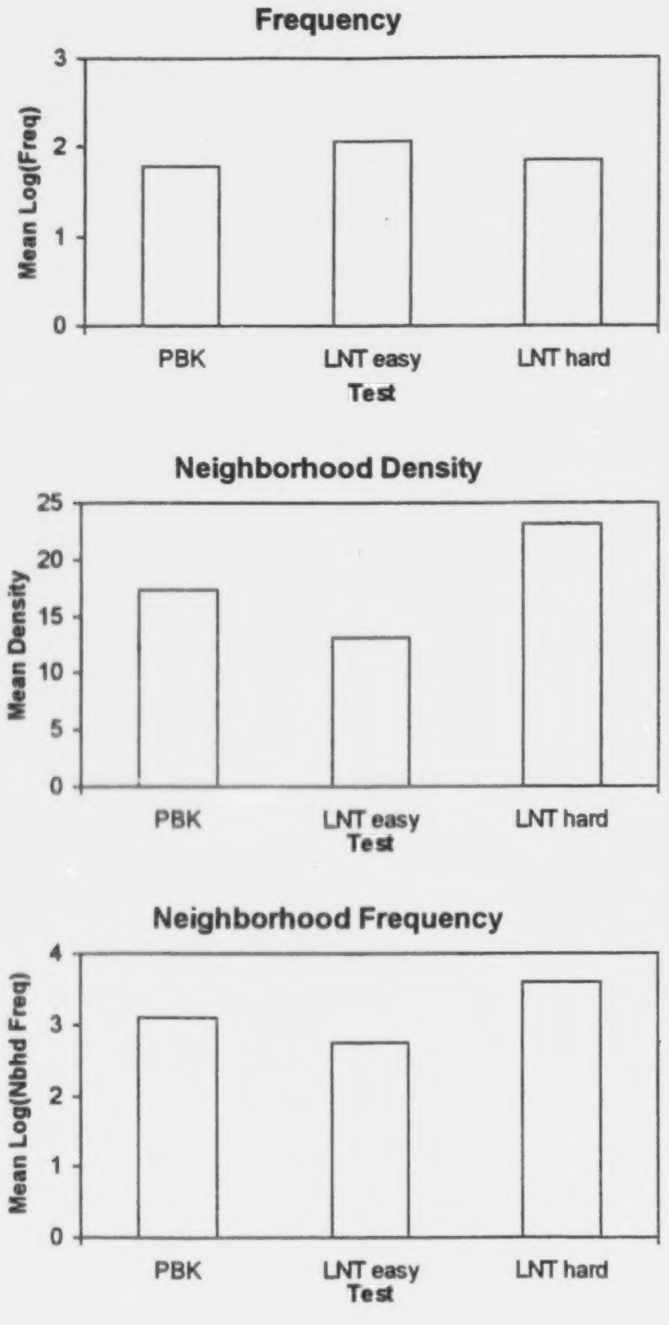


Figure 9. Comparison of mean log frequency, neighborhood density, and log neighborhood frequency of PBK, LNT easy, and LNT hard words.

Finally, note that for both the PBK and the LNT, pediatric cochlear implant users' performance is relatively better when scored by words correct than by phonemes correct. In the case of the PBK, actual scores were equivalent to predicted scores for words correct, but worse for phonemes correct. For the LNT, actual scores were better for words correct, but equivalent for phonemes correct. These differences demonstrate that words are the primary perceptual units of the recognition process in these listeners. The phoneme is a useful theoretical and computational construct only insofar as it represents phonetic evidence for differentiating word candidates, but the word apparently has primacy and special status in the recognition process. This is shown here by the pattern of results for both the PBK and LNT tests when scored separately by phonemes and words correct.

Conclusions

By comparing observed performance to a phoneme confusion model of spoken word recognition, we have successfully replicated a number of results from behavioral studies of pediatric cochlear implant users. These simulations showed that pediatric cochlear implant users are sensitive to the phonetic similarity of percepts to an internalized lexicon of words (Kirk et al. 1995). The phoneme confusion model had no concept of a lexicon, and was thus unable to compensate for incorrect phonemic information which resulted in non-existing words. The phoneme confusion model also revealed differences in the familiarity of words on the PBK and LNT tests. The phoneme confusion model was equally 'unfamiliar' with all words, and predicted equivalent performance on these two tests. In addition, the model predicted no differences between LNT easy and hard word lists. The model also shows that children who perform relatively well on the PBK test (the 'Stars') have made significant progress in developing a mental lexicon, a crucial step in acquiring spoken language (Pisoni et al. this volume). These children performed far better than predicted on the LNT, and at three years post-implant also outperformed the model on the PBK.

In addition, we also found that children who performed poorly on the PBK (the 'Controls') appear to be following the same developmental process as the 'Stars,' but are progressing more slowly over time. Their performance at three years post-implant is approaching the one year post-implant data for the 'Stars.' Like the 'Stars,' it was found that the 'Control' group performed better than expected when words on the LNT were presented, indicating that they have learned some words, and have encoded and stored representations of these words in lexical memory. The performance of the 'Controls' on the PBK was worse than expected, indicating that the 'Controls' have made only limited progress in acquiring a lexicon compared to the 'Stars.' The difference between the 'Stars' and 'Controls' therefore appears to be more quantitative than qualitative, suggesting that time and clinical intervention might significantly improve the language of the poorer performing children by increasing the number of words they have in their lexicons.

References

Black, J.W. (1957). Multiple-choice intelligibility tests. *Journal of Speech and Hearing Disorders*, 22(2): 213-235.

Boothroyd, A. and Nittrouer, S. (1988). Mathematical treatment of context effects in phoneme and word recognition. *Journal of the Acoustical Society of America*, 84(1): 101-114.

Broe, M. (1993). *Specification theory: The treatment of redundancy in generative phonology.* Unpublished Ph.D. dissertation, University of Edinburgh.

Frisch, S. (1996). *Similarity and frequency in phonology*. Unpublished Ph.D. dissertation, Northwestern University.

Haskins, H. A. (1949). *A phonetically balanced test of speech discrimination for children*. Unpublished Master's Thesis, Northwestern University.

Kirk, K. I., Pisoni, D. B., and Osberger, M. (1995). Lexical effects on spoken word recognition by pediatric cochlear implant users. *Ear and Hearing*, 16(5): 470-481.

Goldinger, S. D., Luce, P. A., and Pisoni, D. B. (1989). Priming lexical neighbors of spoken words: Effects of competition and inhibition. *Journal of Memory and Language*, 28, 501-518.

Logan, J. S. (1992). *A computational analysis of young children's lexicons* (Research on Spoken Language Processing Technical Report No. 8). Bloomington, IN: Speech Research Laboratory, Indiana University.

Luce, P. A. (1986). *Neighborhoods of words in the mental lexicon* (Research on Speech Perception Technical Report No. 6). Bloomington, IN: Speech Research Laboratory, Indiana University.

Luce, P. A. and Pisoni, D. B. (1998). Recognizing spoken words: The Neighborhood Activation Model. *Ear and Hearing*, 19, 1-36.

Luce, P. A., Pisoni, D. B., and Goldinger, S. D. (1990). Similarity neighborhoods of spoken words. In G. Altmann (ed.), *Cognitive models of speech processing: Psycholinguistic and computational perspectives*. Cambridge, MA: MIT Press.

Meyer, T. A. and Pisoni, D. B. (this volume). Some computational analyses of the PBK test: Effects of frequency and lexical density on spoken word recognition. In *Research on Spoken Language Processing Progress Report No. 21*. Bloomington, IN: Speech Research Laboratory, Indiana University.

Nittrouer, S. and Boothroyd, A. (1990). Context effects in phoneme and word recognition by young children and older adults. *Journal of the Acoustical Society of America*, 87(6): 2705-2715.

Nusbaum, H. C., Pisoni, D. B., and Davis, C. K. (1984). Sizing up the Hoosier Mental Lexicon: Measuring the familiarity of 20,000 words. In D. Pisoni (ed.), *Research on Speech Perception Progress Report No. 10* (pp. 357-376). Bloomington, IN: Speech Research Laboratory, Indiana University.

Pisoni, D. B., Svirsky, M. A., Kirk, K. I., and Miyamoto, R. T. (this volume). Looking at the "Stars": A first report on the intercorrelations among measures of speech perception, intelligibility, and language in pediatric cochlear implant users. In *Research on Spoken Language Processing Progress Report No. 21*. Bloomington, IN: Speech Research Laboratory, Indiana University.

Rabinowitz, W. M., Eddington, D. K., Delhorne, L. A., and Cuneo, P. A. (1992). Relations among different measures of speech reception in subjects using a cochlear implant. *Journal of the Acoustical Society of America*, 92(4): 1869-1881.

Robbins, A. M., Renshaw, J. J., Miyamoto, R. T., Osberger, M. J., and Pope, M. L. (1988). *Minimal pairs test*. Indianapolis, IN: Indiana University School of Medicine.

Sommers, M. S., Kirk, K. I., and Pisoni, D. B. (1997). Some considerations in evaluating spoken word recognition by normal-hearing, noise-masked normal hearing, and cochlear implant listeners I: The effects of response format. *Ear and Hearing*, 18: 89-99.

Stanley, R. (1967). Redundancy rules in phonology. *Language*, 43: 393-436.

Appendix

The computational analysis in this paper uses the phoneme inventory in the on-line version of Webster's pocket dictionary (Nusbaum, Pisoni, and Davis 1984), known as the 'Hoosier Mental Lexicon' or HML. Phonemes are indicated here with both IPA transcription and the equivalent HML character. Features used for the consonants and vowels, along with the broad categories of contrast by which they were grouped, are given in (A1) and (A2).

(A1) Features used to represent the English consonant inventory:

a. ARTICULATOR features (Place category):

IPA	p	b	f	v	m	t	d	θ	ð	s	z	ʃ	ʒ	tʃ	dʒ	k	g	ŋ	l	r	n	w	y	h
HML	p	b	f	v	m	t	d	T	D	s	z	S	Z	C	J	k	g	G	l	r	n	w	y	h
labial	+	+	+	+	+															+		+		
coronal						+	+	+	+	+	+	+	+	+	+				+	+	+		+	
dorsal																+	+	+						

b. PLACE features (Place category):

IPA	p	b	f	v	m	t	d	θ	ð	s	z	ʃ	ʒ	tʃ	dʒ	k	g	ŋ	l	r	n	w	y	h
HML	p	b	f	v	m	t	d	T	D	s	z	S	Z	C	J	k	g	G	l	r	n	w	y	h
bilabial	+	+			+															+		+		
dental			+	+				+	+															
alveolar						+	+			+	+								+	+	+			
palatal												+	+	+	+								+	
velar																+	+	+						

c. STRICTURE features (Manner category):

IPA	p	b	f	v	m	t	d	θ	ð	s	z	ʃ	ʒ	tʃ	dʒ	k	g	ŋ	l	r	n	w	y	h
HML	p	b	f	v	m	t	d	T	D	s	z	S	Z	C	J	k	g	G	l	r	n	w	y	h
obstruent	+	+	+	+		+	+	+	+	+	+	+	+	+	+	+	+							+
sonorant					+													+	+	+	+	+	+	
stop	+	+			+	+	+							+	+	+	+	+						
continuant			+	+				+	+	+	+	+	+											+
glide																						+	+	
consonantal					+													+	+	+	+			

d. MANNER features (Manner category):

IPA	p	b	f	v	m	t	d	θ	ð	s	z	ʃ	ʒ	tʃ	dʒ	k	g	ŋ	l	r	n	w	y	h
HML	p	b	f	v	m	t	d	T	D	s	z	S	Z	C	J	k	g	G	l	r	n	w	y	h
oral	+	+				+	+									+	+							
affricate														+	+									
strident										+	+	+	+	+	+									
distributed			+	+				+	+															+
lateral																			+					
rhotic																				+				
nasal					+													+			+			

e. LARYNGEAL features (Voicing category):

IPA	p	b	f	v	m	t	d	θ	ð	s	z	ʃ	ʒ	tʃ	dʒ	k	g	ŋ	l	r	n	w	y	h
HML	p	b	f	v	m	t	d	T	D	s	z	S	Z	C	J	k	g	G	l	r	n	w	y	h
voice		+		+	+		+		+		+		+		+		+	+	+	+	+	+	+	
voiceless	+		+			+		+		+		+		+		+								+
spread glottis																								+

(A2) Features used to represent the English vowel inventory:

a. VOWEL PLACE features:

IPA	i	ɪ	e	ɛ	æ	a	ʌ	u	ʊ	o	ɔ	oᶦ	aᶦ	ɔᵘ	ə	ə̣	ɨ̣	ɚ	ḷ	ṇ	ṃ
HML	i	I	e	E	@	a	^	u	U	o	c	O	Y	W	x	|	X	R	L	N	M
front	+	+	+	+	+	+						+	+			+				+	
mid							+								+			+	+		
back								+	+	+	+	+		+							
labial								+	+	+	+	+		+							+

b. VOWEL HEIGHT features:

IPA	i	ɪ	e	ɛ	æ	a	ʌ	u	ʊ	o	ɔ	oᶦ	aᶦ	ɔᵘ	ə	ə̣	ɨ̣	ɚ	ḷ	ṇ	ṃ
HML	i	I	e	E	@	a	^	u	U	o	c	O	Y	W	x	|	X	R	L	N	M
high	+	+						+	+			+	+	+			+		+		
mid-high			+							+		+			+	+		+			
mid-low				+			+				+			+							
low					+	+							+								

c. VOWEL MANNER features:

IPA	i	ɪ	e	ɛ	æ	a	ʌ	u	ʊ	o	ɔ	oᶦ	aᶦ	ɔᵘ	ə	ə̣	ɨ̣	ɚ	ḷ	ṇ	ṃ
HML	i	I	e	E	@	a	^	u	U	o	c	O	Y	W	x	|	X	R	L	N	M
static	+	+	+	+	+	+	+	+	+	+	+										
dynamic												+	+	+							
tense	+		+			+	+	+		+	+	+	+	+							
lax		+		+	+				+												
stressed	+	+	+	+	+	+	+	+	+	+	+	+	+	+							
unstressed															+	+	+				
consonantal																		+	+	+	+

Lexical Competition in Spoken English Words[1]

Shigeaki Amano[2]

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405

---

[1] This work is supported in part by NIH-NIDCD Research Grant DC00111 to Indiana University. I am grateful to David Pisoni and Gina Torretta who allowed me to use their database. They also gave me valuable comments and suggestions on an earlier draft of this paper.

[2] NTT Basic Research Laboratories, 3-1 Morinosato-Wakamiya, Atsugi, Kanagawa, 243-01 Japan. E-mail: amano@av-hp.brl.ntt.co.jp

Lexical Competition in Spoken English Words

Abstract. The "rime cognate," which contains the same phoneme sequence as the rime of a given word, is proposed as a new lexical competitor set for spoken word recognition. Partial correlation analyses showed that the rime cognate has a greater negative contribution to the identification of English spoken words than the lexical neighborhood and the word-initial cohort which had been commonly used as a lexical competitor set. The analyses also showed that the rime cognate includes the reliable part of the neighborhood and the cohort, which indicates that the important parts of these two different competitor sets are integrated into the rime cognate.

Introduction

Research on spoken word recognition suggests that multiple word candidates are activated during the recognition process, and that these candidates compete with each other. Some empirical support for activation of multiple word candidates has been obtained in a cross-modal semantic priming task. For instance, Shillcock (1990) has shown that, when a carrier word is processed, an embedded word in the carrier word is also activated. For example, "bone" is activated when "trombone" is processed. Connine, Blasko, and Wang (1994) also used the cross-modal semantic priming task to investigate spoken word recognition. They found that multiple word candidates are activated when a spoken word has an ambiguous initial phoneme. Zwitserlood (1989) has also shown that sentential-semantic contexts are used for selecting one of the activated word candidates.

Although these studies indicate that the multiple word candidates are activated during the recognition process, there is further support that the multiple word candidates compete with each other. For example, McQueen, Norris, and Cutler (1994) measured reaction times for detecting the monosyllabic target word which is embedded in the second syllable in a carrier nonword with a weak-strong syllable pattern. They found that the reaction times for the target word are longer when the carrier nonword corresponded to the beginning of an other multisyllabic word than when it did not. For example, detection of "mess" in "domes" (which is a part of "domestic") required longer reaction time than detection of "mess" in "nemess," indicating that "domestic" competes with "mess." Their results suggest that the multisyllabic word which shares its beginning part with the carrier nonword competes the monosyllabic target word.

Evidence for such lexical competition has also been obtained in the studies of phonological priming (Hamburger & Slowiaczek, 1996, Slowiaczek & Hamburger, 1992). Slowiaczek and Hamburger (1992) have found interference effects on reaction time for shadowing of the target word when the prime has an initial three phoneme overlap with the target word, suggesting the presence of competition between the target word and the prime.

Other evidence of lexical competition has been reported by Vroomen and de Gelder (1995). They found that facilitative cross-modal priming effects are larger for auditory primes with few or no competitors than for auditory primes with many competitors. Their results suggest that competitors have an inhibitory effect on word recognition according to their numbers. They also found that the difference of priming effects according to the number of competitors disappeared when there was no interstimulus interval between auditory prime and visual target word. This result suggests that lexical competition requires some amount of time to be effective.

Several recent models of spoken word recognition explicitly assume competition among word candidates. For example, the TRACE model (McClelland & Elman, 1986) and the SHORTLIST model (Norris, 1994; Norris, McQueen, & Cutler, 1995) both incorporate lexical competition by adopting inhibitory connections among word candidates. On the other hand, other models such as the original cohort model (Marslen-Wilson & Welsh, 1978), the new cohort model (Marslen-Wilson, 1987), and the Neighborhood Activation Model (Luce, 1986) do not have such inhibitory relationship among word candidates. However, as pointed by Frauenfelder (1996), these models incorporate the competition implicitly, because they use a decision rule which determines the "winning" candidates by taking account of the status of all other word candidates.

The characteristics of the word candidate set have been described by three variables in previous studies. The first variable is "density" which is the number of words in a candidate set (e.g., Luce, 1986; Frauenfelder, Baayen, Hellwig, & Schreuder, 1993). The second variable is "mean frequency" which is the averaged frequency of words in a candidate set (e.g., Frauenfelder, Baayen, Hellwig, & Schreuder, 1993; Luce, 1986; Marslen-Wilson, 1990). The third variable is "maximum frequency" which is the highest frequency of a word in a candidate set (Bard, 1990; Bard & Shillcock, 1993).

Although these variables has been used commonly, there are some contradictions about which variable has the largest contribution to word recognition (e.g., Bard, 1990; Bard & Shillcock, 1993). More importantly, there has been little general agreement on the set of word candidates. That is, there is a discrepancy on what kind of words are included in the set. Previous studies assume at least two different candidate sets for spoken word recognition; the lexical neighborhood and the word-initial cohort.

A lexical neighborhood is defined as a collection of words which have single phoneme substitution with a target word (e.g., Frauenfelder, 1990; Frauenfelder, Baayen, Hellwig, & Schreuder, 1993; Pisoni, Nusbaum, Luce & Slowiaczek, 1985). Other definition for the neighborhood includes words with single phoneme deletion or addition to the target words in addition to the substitution (e.g., Goldinger, 1989; Luce, 1986; Sommers, 1996). The former definition is used as the neighborhood in this paper, because it is simpler than the latter, and no substantial differences were found in the statistical characteristics between the two definitions (Frauenfelder, Baayen, Hellwig, & Schreuder, 1993).

Several recent studies indicate that the neighborhoods have significant effects on word recognition performance. For example, neighborhood density and frequency negatively correlate with recognition rate of a target word (Luce, 1986). Neighborhood density has inhibitory effects on reaction times in lexical decision and naming for a target word (Goldinger, 1989). Low frequency words in a neighborhood have negative priming effects on target word recognition (Goldinger, Luce, & Pisoni, 1989). Recognition of two-syllable target words (spondees) is affected by neighborhood characteristics of each syllable in the word (Cluff & Luce, 1990). Older adults have difficulty recognizing a target word if density and frequency is high in neighborhood (Sommers, 1996). Taken together, all these studies clearly indicate that the lexical neighborhood affects its recognition.

In contrast to the role of the neighborhood, several studies have shown that reliable word candidate set is the cohort which is a set of words sharing the initial part of phoneme sequences with a target word (e.g., Marslen-Wilson & Welsh, 1978; Marslen-Wilson, 1989; Bard & Shillcock, 1993). For example, Marslen-Wilson (1990) found that the isolation point (IP) of a word in the gating task (Grosjean, 1980) is later for a word with high frequency competitors in its cohort than it is for a word with low frequency competitors. His results show that word candidates in the cohort affect the word recognition and that the amount of effect is a function of their frequency.

In the definition of the neighborhood and the cohort, an English word is treated as a simple sequence of phonemes. However, some linguistic studies argue that words can be sub-divided into syllables. The syllable, in tern, has two main parts, an onset and a rime (Cairns & Feinstein, 1982; Fudge, 1969; Halle & Vergnaud, 1980). The onset contains the initial consonant cluster of the syllable. The rime contains the peak (vowel nucleus) and the coda (final consonant cluster).

Evidence for the division of the syllable into onset and rime components has been obtained in numerous psycholinguistic studies. For example, support has been found in the studies of speech errors (MacKay, 1972; Stemberger & Treiman, 1986), short-term memory of syllables (Treiman & Danis, 1988), word games for syllable blending (Treiman, 1983, 1986), and for syllable dividing and transforming (Treiman, 1986). Further support has been obtained from children's performance of word games for syllable transforming (Treiman, 1985), rime identification (Lenel & Cantor, 1981; Treiman & Zukowski, 1996), and word categorization (Kirtley, Bryant, MacLean, & Bradley, 1989).

Given the psycholinguistic importance of the onset and the rime, two alternatives can be considered as word candidates for lexical competition. One is the set of words which share the onset, the other is the set of words which are related to each other through the rime.

Broadly speaking, the onset-sharing set corresponds to the word-initial cohort. When a word begins with a consonant, the cohort which shares the initial part of word coincides with the onset-sharing set. When a word begins with a vowel, the cohort does not coincides with the onset-sharing set. However, words beginning with a consonant are more frequent than the words beginning with a vowel in English. 79.7% of words begin with a consonant when a search is conducted using a computerized dictionary (Nusbaum, Pisoni, & Davis, 1984) which contains 19,295 words. This fact means that the cohort roughly corresponds to the onset-sharing set. Therefore, it is not necessary to consider it here.

On the other hand, the set of words which are related to each other through the rime has not been investigated. This set is hereafter called a "rime cognate." The precise definition of the rime cognate is a set of the words which contain the same sequence of phonemes with the rime of a given word. The lexical matching procedure used to produce the rime cognate is hereafter called rime matching strategy (RMS).

For example, if a word "cat" is given, its rime cognate produced by the RMS contains, for instance, "hat", "flat", and "sprat" because they share "at" with the rime of the "cat." The rime cognate also contains multisyllabic words such as "matins" having "at" in the first syllable, and "cravat" having "at" in the second syllable.

In the present study, the three sets of word candidates, the neighborhood, the cohort, and the rime cognate, are compared for their reliability as a lexical competitor set in spoken word recognition. Three sets of analyses were carried out. The first analysis showed that the rime cognate is better than the neighborhood and the cohort in terms of correlations with word recognition rates. The second and third analyses showed that the rime cognate includes the reliable part of the neighborhood and the cohort, which indicate that the rime cognate integrates the useful portions of the neighborhood and the cohort.

Analysis 1

This analysis addresses the question of which set of word candidates is the best predictor for lexical competition in spoken word recognition among the neighborhood, cohort, and rime cognate. The best candidate set should have the greatest negative correlations to word recognition rate in terms of the density, the mean frequency, and/or the maximum frequency, as a reflections of lexical competition.

Materials

Analysis 1 was conducted on the recognition rates of English spoken words which were extracted from the "Easy-Hard" word multi-talker speech database (Torretta, 1995). The database contains recognition rates of 150 monosyllabic words pronounced by 10 talkers at three speaking rates of slow, medium, and fast. The 150 words pronounced by each talker at each speaking rate were presented to 10 subjects to obtain a percent correct identification score. A total of 300 subjects (10 subjects X 10 talkers X 3 rates) participated in supplying responses for the database.

Three items out of the 150 words were not used for the analysis. They were "white" and two "wrong"s. It is because the "white" might be a CCVC word rather than a CVC word, and because there are two entries for the "wrong" in the database. After excluding these three items, the recognition rates of the 147 CVC words (See Appendix) in three speaking rates were used for Analysis 1. There were 441 items in total. Word recognition rates for these items were obtained by averaging the identification score over ten talkers.

Definition of Candidate Set

Definitions of the candidate sets used for Analysis 1 are shown in Table 1. "Neighborhood" is defined as a union of three set of words of the first, the second, and the third phoneme substitution with a target word (Pisoni, Nusbaum, Luce, & Slowiaczek, 1985; Frauenfelder, 1990; Frauenfelder, Baayen, Hellwig, & Schreuder, 1993).

"Cohort" is defined as a set of words which share the first two phonemes (i.e., CV) with a target CVC word. This definition is consistent with the word-initial cohort which was used in the studies of Bard and Shillcock (1993) and Marslen-Wilson, Moss, and van Halen (1996).

"Rime Cognate" is defined as a set of words which have the same phoneme sequence with the rime (i.e., VC) of a target CVC word. The rime cognate used in Analysis 1 does not include the words which have the same phoneme sequence with the rime of the target CVC word at the second syllable or at the later syllables. For example, the rime cognate of the target word of "cat" does not include "cravat" which has the same phoneme sequence with the rime at the second syllable. This was done because it is hard to imagine that such word candidates are activated for a target word presented in isolation. Because there is enough silence before the target word, it clearly indicates the beginning of the target word and it removes the possibilities for the activation of such word candidates. However, if the target word is embedded in continuous speech, it would be reasonable to include such word candidates, because the beginning of the word is sometimes ambiguous in the continuous speech.

Table 1.

Phoneme sequence pattern of word candidates in neighborhood, cohort, and rime cognate for CVC target word.

Phoneme Sequence Pattern

Neighborhood	?VC C?C CV?
Cohort	CV(*)
Rime Cognate	(c+)VC(*)

Note.

C = Common consonant to target word,
V = Common vowel to target word,
? = Any one phoneme,
* = Any length of phoneme sequence,
c+ = Any length of consonant sequence.
Items in parenthesis can be null.

Method

Density, mean log frequency, and the maximum log frequency were obtained for the neighborhood, the cohort, and the rime cognate of the 147 target words using a computerized dictionary (Nusbaum, Pisoni, & Davis, 1984) which contains 19,295 words with Kucera and Francis (1967) word count. Log frequency of the 147 target words was also obtained from the dictionary. Summary statistics for these variables are shown in Table 2.

By excluding the factor of log frequency of the target word, Spearman's rank order partial correlation coefficient was calculated between the word recognition rate and the three variables of density, mean log frequency, and maximum log frequency. The Spearman's rank order partial correlation coefficient is suitable to the analysis because there is a weak tendency that the density and the frequency of the neighborhood are higher for high-frequency target words than for low-frequency target words (Frauenfelder, Baayen, Hellwig, & Schreuder, 1993; Pisoni, Nusbaum, Luce, & Slowiaczek, 1985), and because the distributions are skewed for the density, mean log frequency and maximum log frequency of the neighborhood (Frauenfelder, Baayen, Hellwig, & Schreuder, 1993) and cohort (Bard & Shillcock, 1993) as well as for the log frequency of a target word.

Table 2.

Statistics of the CVC words for Analysis 1 (N=147).

	Mean	SD	Minimum	Maximum
Target Word				
Log Frequency	1.39	0.83	0.30	3.99
Neighborhood				
Density	17.56	6.77	1	31
Mean Log Frequency	1.84	0.48	0.56	2.89
Maximum Log Frequency	2.69	0.70	0.85	4.03
Cohort				
Density	56.27	61.80	1	325
Mean Log Frequency	1.34	0.48	0.30	2.89
Maximum Log Frequency	2.43	0.79	0.30	4.42
Rime Cognate				
Density	109.00	143.60	1	973
Mean Log Frequency	1.34	0.37	0.30	2.27
Maximum Log Frequency	2.64	0.77	0.30	4.46

Results

The Spearman's rank order partial correlation coefficients are shown in Figure 1. The results are shown for the data of all speaking rates in Figure 1, because no substantial differences were observed among the results of the slow, medium, and fast speaking rate.

The partial correlation coefficient significantly differed from zero for all variables of the rime cognate; density ($r = -.293$, $p < .01$), mean log frequency ($r = -.145$, $p < .01$), and maximum frequency ($r = -.215$, $p < .01$). The partial correlation coefficient also differed from zero for the density of the neighborhood ($r = -.192$, $p < .01$), and the density of the cohort ($r = -.104$, $p < .05$). However, it was not significantly different from zero for the mean log frequency and the maximum frequency of the neighborhood and cohort.

For the density, the partial correlation coefficient was different between the rime cognate and the neighborhood ($p < .05$), and between the rime cognate and the cohort ($p < .01$). For the mean log frequency, the partial correlation coefficient was different between the rime cognate and the cohort ($p < .01$), and between the cohort and the neighborhood ($p < .05$). For the maximum log frequency, the partial correlation coefficient was different between the rime cognate and the neighborhood ($p < .05$), and between the rime cognate and the cohort ($p < .01$).

For the rime cognate, the partial correlation coefficient was different between the density and the mean log frequency ($p < .05$), and between the mean log frequency and the maximum log frequency ($p < .01$). However, it was not significantly different between the density and the maximum log frequency. For the neighborhood, the partial correlation coefficient was different between the density and the mean log frequency ($p < .05$), and between the density and the maximum log frequency ($p < .05$). For the cohort, the partial correlation coefficient was different between the density and the mean log frequency ($p < .05$), and between the density and the maximum log frequency ($p < .05$).

Insert Figure 1 about here.

Discussion

Results of this first analysis clearly show that the rime cognate is a better predictor of lexical competition in spoken word recognition than the neighborhood and the cohort. The partial correlation coefficient has a larger negative value for the rime cognate than the neighborhood and the cohort in terms of the density, the mean log frequency, and the maximum log frequency.

All partial correlation coefficients were negative for the rime cognate. This pattern indicates that the set of words which have the same phoneme sequence with the rime (i.e., VC) of a CVC target word are activated and they compete against the target word. Notice that the negative effects are not an artifact of the target frequency, because the contribution of the target frequency is excluded by calculating partial correlations. Hence, the effects are purely based on the structural characteristics of the rime cognate.

For the rime cognate, the density had the greatest contribution to word recognition, the maximum log frequency was the second, and the mean log frequency was the least. Although there were significant differences between the density and the mean log frequency, there were no significant differences between the density and the maximum log frequency. This fact is ambiguous in terms of the number of activated competitors. Significant partial correlation coefficient for the density indicates that all of the competitors in the rime cognate affect the word recognition. On the other hand, significant partial correlation coefficient for the maximum log frequency indicates that only one competitor with the highest frequency in the rime cognate affects the word recognition.

The significant partial correlation coefficient for the maximum log frequency might be an artifact of the effect of mean log frequency. If the maximum log frequency has a high positive correlation with the mean log frequency, it automatically produces significant relationship to word recognition when the mean log frequency is significantly related to word recognition. And, this appears to be the case for the current data set. Spearman's rank order partial correlation coefficient was very high between the mean log frequency and the maximum log frequency ($r = .844$, $p < .001$). Therefore, the pattern is consistent. All the competitors in the rime cognate affect the word recognition, but one competitor with the highest frequency does not.

How does both the density and the frequency of the rime cognate affect word recognition? One possibility is that the amount of effectiveness of each competitor is modulated by its frequency. That is, a high frequency competitor is more effective than low frequency competitor in the rime cognate.

Bard (1990) and Bard and Shillcock (1993) have already mentioned this possibility of frequency modulation of competitors. And, it has been incorporated in some models of spoken word recognition. For example, TRACE model (McClelland & Elman, 1986) achieves such modulation by changing the resting activation level of a word unit.

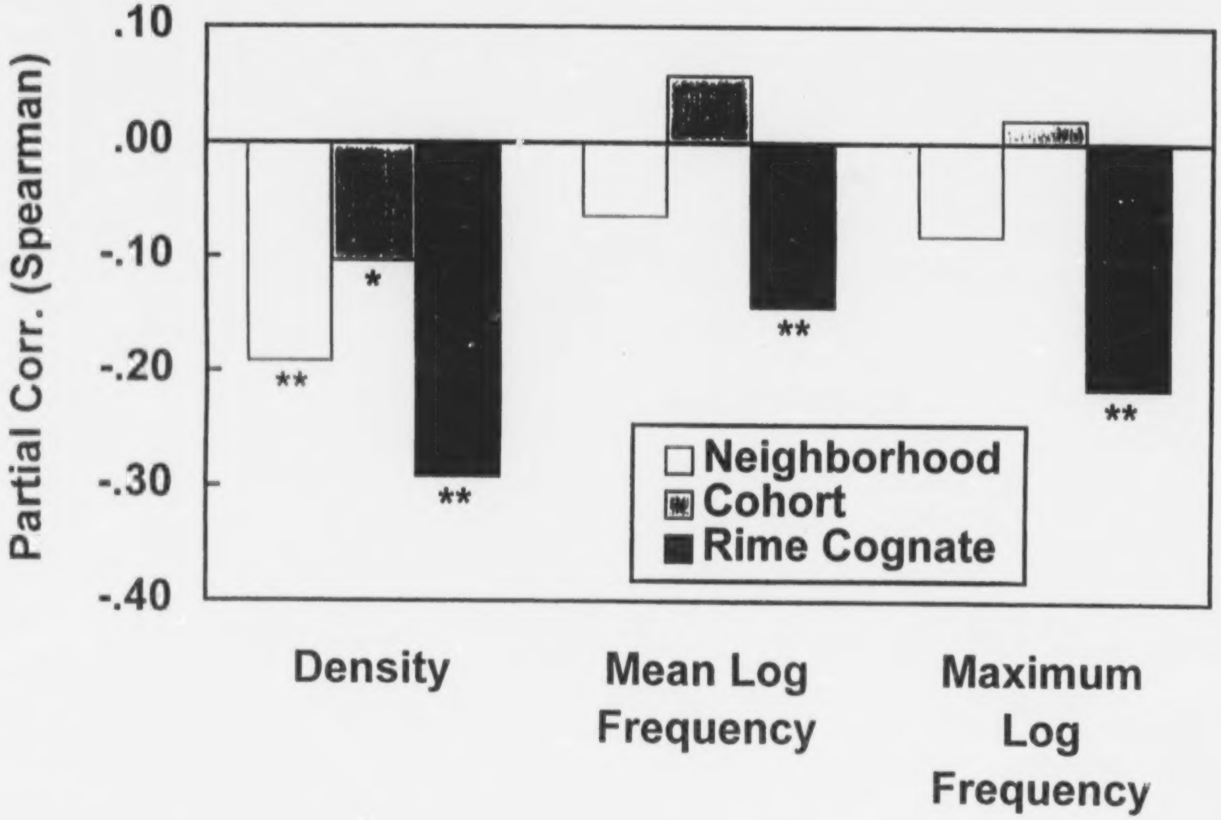


Figure 1. Spearman's rank order partial correlation coefficient between word recognition rate and density, mean log frequency, and maximum log frequency of neighborhood, cohort, and rime cognate. Excluding factor for the partial correlation is log frequency of target word. Single and double asterisks respectively represent difference from zero at 5% and 1% significance level.

However, Bard (1990) and Bard and Shillcock (1993) analyzed the cohort only. And, the TRACE model basically uses the cohort-type candidate set. Further research, therefore, is necessary to reveal the characteristics of the frequency modulation in the rime cognate.

Analysis 2

Although Analysis 1 showed that the rime cognate is reliable on spoken word recognition in terms of partial correlation with identification performance, many studies have shown that the neighborhood is also reliable. Analysis 2 addresses the question of why the neighborhood is reliable in some degree.

The conventional view of a lexical neighborhood is a union of positional neighborhoods in which a phoneme is substituted to a target word at one of phoneme positions (e.g., Frauenfelder, 1990; Frauenfelder, Baayen, Hellwig, & Schreuder, 1993; Pisoni, Nusbaum, Luce & Slowiaczek, 1985). In case of the CVC words, the conventional neighborhood consists of the positional neighborhoods at the first (?VC), the second (C?C), and the third phoneme position (?VC) as shown in Table 1. Among these positional neighborhoods, the positional neighborhood with the first phoneme substitution (?VC) is included as a part of the rime cognate ((c+)VC(*)), because it shares the second and the third phoneme (i.e., the rime) with the target word as the rime cognate does.

It is possible that the conventional neighborhood is reliable because this positional neighborhood contributes much to the word recognition. More specifically, the hypothesis is that a positional neighborhood with the first phoneme substitution would be more reliable than other positional neighborhoods with the second or the third phoneme substitution. Analysis 2 assesses this hypothesis.

Materials

Word recognition rates used in Analysis 1 were also used in Analysis 2.

Definition of Candidate Set

Analysis 2 was conducted on the positional neighborhoods of which a phoneme was substituted to the target word at one of phoneme positions. The positional neighborhood of the first phoneme position is the set of words which share the second phoneme and the third phoneme with the target word but not the first phoneme (?VC). The positional neighborhood of the second phoneme position is the set of words which share the first phoneme and the third phoneme with the target word but not the second phoneme (C?C). The positional neighborhood of the third phoneme position is the set of words which share the first phoneme and the second phoneme with the target word but not the third phoneme (CV?).

Method

For each positional neighborhood of the 147 target words, density, mean log frequency, and the maximum log frequency were obtained from the same computerized dictionary as Analysis 1. Summary statistics for these variables are shown in Table 3.

By excluding the factor of log frequency of a target word, Spearman's rank order partial correlation coefficient was calculated between word recognition rates and the three variables of density, mean log frequency, and maximum log frequency for each neighborhood.

Table 3.

Statistics of the positional neighborhood of the CVC words for Analysis 2 (N=147).

Positional Neighborhood	Mean	SD	Minimum	Maximum
1st Phoneme Position				
Density	7.03	3.86	0	16
Mean Log Frequency	1.59	0.73	2.97	0.39
Maximum Log Frequency	2.13	0.96	4.03	0.63
2nd Phoneme Position				
Density	4.76	2.82	0	10
Mean Log Frequency	1.49	0.74	2.92	0.46
Maximum Log Frequency	1.89	0.95	3.71	0.72
3rd Phoneme Position				
Density	5.78	2.94	0	13
Mean Log Frequency	1.52	0.71	3.09	0.44
Maximum Log Frequency	1.99	0.89	3.86	0.67

Results

Figure 2 shows Spearman's rank order partial correlation coefficients between word recognition rate and the variables of the positional neighborhoods. Figure 2 also shows the partial correlation obtained in Analysis 1 for the conventional neighborhood which is represented by "U" (for "union").

For the positional neighborhood of the first phoneme position (?VC), the partial correlation coefficient significantly differed from zero for the density ($r = -.158$, $p < .01$), the mean log frequency ($r = -.213$, $p < .01$), and the maximum log frequency ($r = -.216$, $p < .01$). For the positional neighborhood of the second phoneme position (C?C), the partial correlation coefficient significantly differed from zero only for the density ($r = -.155$, $p < .01$). For the conventional neighborhood (U), the partial correlation coefficient significantly differed from zero for the density ($r = -.192$, $p < .01$) as shown in Analysis 1. No significant differences were observed between any pair of these significant partial correlation coefficients.

For the density, the partial correlation coefficient was significantly different between the conventional neighborhood (U) and the positional neighborhood of the third phoneme position (CV?) ($p < .01$).

For the mean log frequency, the partial correlation coefficient for the positional neighborhood of the first phoneme position (?VC) was significantly different from that for the conventional neighborhood (U) ($p < .01$), for the positional neighborhood of the second phoneme position (C?C) ($p < .01$), and for the positional neighborhood of the third phoneme position (CV?) ($p < .05$).

For the maximum log frequency, the partial correlation coefficient for the positional neighborhood of the first phoneme position (?VC) was significantly different from that for the conventional neighborhood (U) ($p < .01$), for the positional neighborhood of the second phoneme position (C?C) ($p < .01$), and for the positional neighborhood of the third phoneme position (CV?) ($p < .05$).

Insert Figure 2 about here.

Discussion

Although the positional neighborhood of the first phoneme substitution (?VC) was not different for the positional neighborhood of the second phoneme substitution (C?C) in terms of the density, it showed a much larger negative partial correlation coefficient than other positional neighborhoods in terms of the density, the mean log frequency, and the maximum log frequency. These results indicate that the positional neighborhood with the first phoneme substitution (?VC) is more reliable than other positional neighborhoods with the second (C?C) or the third phoneme substitution (CV?).

The positional neighborhood of the first phoneme substitution (?VC) is a part of the conventional neighborhood (U) which has reliable effects of density as shown in Analysis 1. Spearman's rank order partial correlation coefficients between these two neighborhoods was positive and high (r = .691) in terms of the density. This pattern means that the conventional neighborhood is reliable because it contains the positional neighborhood of the first phoneme substitution which is reliable.

The positional neighborhood of the first phoneme substitution is included in the rime cognate which is the best candidate set in Analysis 1. Therefore, the results of Analysis 2 suggest that the conventional neighborhood is reliable in some degree because it shares a part with the rime cognate which is more appropriate as a word candidate set.

Analysis 3

In Analysis 1, the cohort was defined as a set of words which share the first two phonemes (i.e., CV part) with a target CVC word. However, cohort can be defined differently because the number of shared phonemes is not explicitly specified for the word-initial cohort. The word-initial cohort can share one or two "segments" at word onset (Marslen-Wilson, 1989) or initial "sequence" (Marslen-Wilson, 1984) which correspond to 100 to 150 ms of speech waveform (Marslen-Wilson, 1987). Therefore, the cohort can be defined as a set of words sharing only an initial phoneme, first two phonemes, or first three phonemes (Marslen-Wilson & Welsh, 1978). These three kind of sets were used in Analysis 3.

In case of the CVC words, the cohort sharing the first three phonemes (i.e., all phonemes) with the target word is completely included in the rime cognate. On the other hand, the cohorts sharing the first phoneme or the first two phonemes are only partially included. Notice that the candidates which are common with the rime cognate are the same across the these cohorts, because the cohorts sharing the first phoneme and the first two phonemes include the cohort sharing the first three phonemes.

However, the proportion of the common candidates with the rime cognate is larger in the cohort sharing the first three phonemes than in other cohorts. In other words, the cohorts sharing the first phoneme or the first two phonemes have larger numbers of uncommon candidates with the rime cognate than the cohort sharing the first three phonemes.

Therefore, it is likely that the cohort sharing the first three phonemes might be more reliable in word recognition than the other cohorts, because the former cohort is a part of the rime cognate which is reliable in the word recognition. This possibility was examined in Analysis 3.

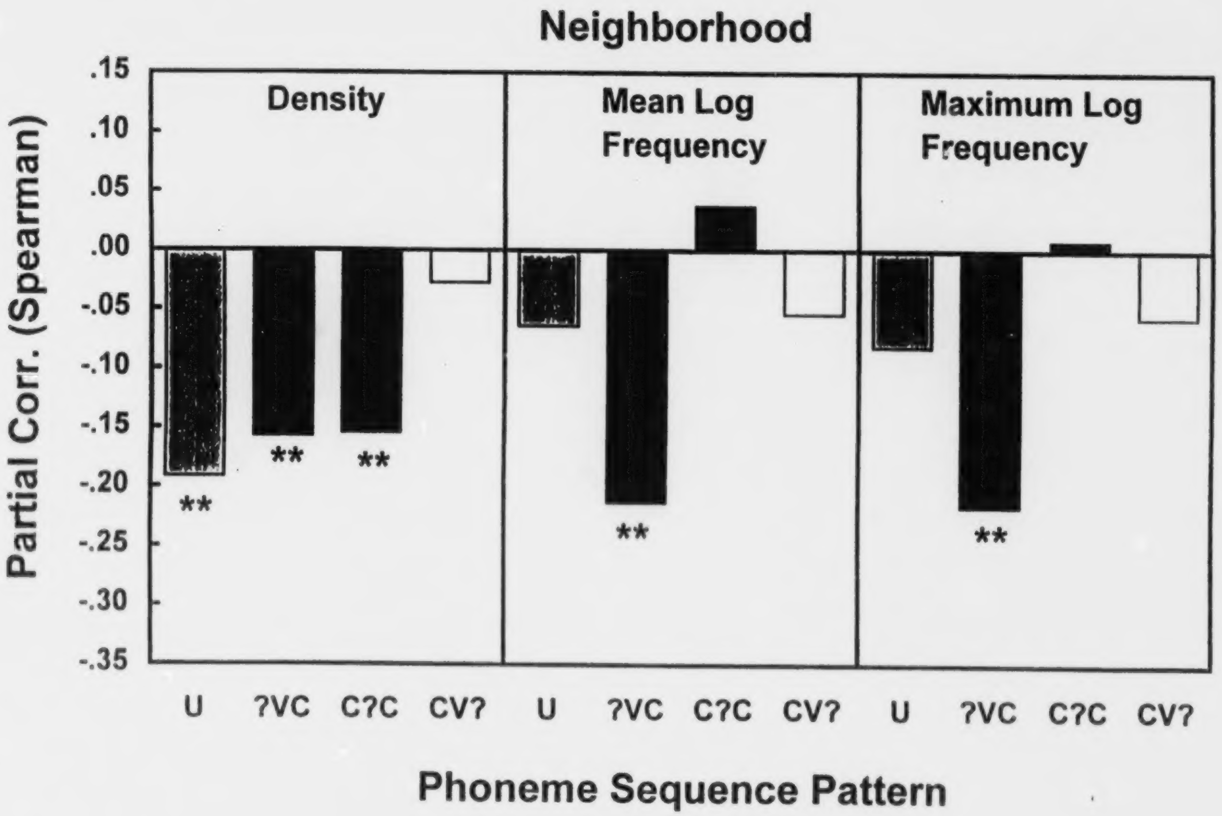


Figure 2. Spearman's rank order partial correlation coefficient between word recognition rate and density, mean log frequency, and maximum log frequency of neighborhood. Excluding factor for the partial correlation is log frequency of target word. "?" represents a substitutable single phoneme. "C" and "V" respectively represent a common consonant and vowel with the CVC target word. "U" represents the conventional neighborhood. Double asterisks show the significant difference from zero ($p < .01$).

Materials

Word recognition rates used in Analysis 1 were also used in Analysis 3.

Definition of Candidate Set

Three cohorts were used in Analysis 3. They were the cohorts sharing the first phoneme, the first two phonemes, and the first three phonemes. The cohort sharing the first phoneme is represented as C(*) in which the common first phoneme with a target word is followed by any sequence of phonemes. The cohort sharing the first two phonemes is represented as CV(*), which is the same one as in Analysis 1. The cohort sharing the first three phonemes is represented as CVC(*) in which an entire phoneme sequence of the target word is followed by any sequence of phonemes.

Method

For each cohort of the 147 target words, density, mean log frequency, and the maximum log frequency were obtained from the same computerized dictionary as Analysis 1. Statistics of these variables are shown in Table 4.

By excluding the factor of log frequency of a target word, Spearman's rank order partial correlation coefficient was calculated between the word recognition rate and the three variables of density, mean log frequency, and maximum log frequency.

Table 4.

Statistics of the cohort of the CVC words for Analysis 3 (N=147).

Cohort	Mean	SD	Minimum	Maximum
One Phoneme Sharing (C(*))				
Density	902.50	520.07	116	1895
Mean Log Frequency	1.46	0.27	1.00	2.09
Maximum Log Frequency	3.36	0.46	2.55	4.42
Two Phoneme Sharing (CV(*))				
Density	56.27	61.80	1	325
Mean Log Frequency	1.34	0.48	0.30	2.89
Maximum Log Frequency	2.43	0.79	0.30	4.42
Three Phoneme Sharing (CVC(*))				
Density	4.65	10.92	0	125
Mean Log Frequency	0.61	0.53	0	2.64
Maximum Log Frequency	0.83	0.73	0	2.64

Results

Figure 3 shows Spearman's rank order partial correlation coefficient. The results of Analysis 1 for the cohort sharing the first two phonemes were replotted as the notation with CV(*) in Figure 3.

For the cohort sharing the first three phonemes (CVC(*)), the partial correlation coefficient significantly differed from zero for the density ($r = -.232$, $p < .01$), the mean log frequency ($r = -.147$, $p < .01$), and the maximum log frequency ($r = -.172$, $p < .01$). For the cohort sharing the first two phonemes (CV(*)), the partial correlation coefficient significantly differed from zero only for the density ($r = -.104$, $p < .05$), as already shown in Analysis 1.

For the cohort sharing the first three phonemes (CVC(*)), the partial correlation coefficient was significantly different between the density and the mean log frequency ($p < .05$), and between the mean log frequency and the maximum log frequency ($p < .01$). But there was no difference of the partial correlation coefficient between the density and the maximum log frequency.

For the density, the partial correlation coefficient for the cohort sharing the first three phonemes (CVC(*)) was significantly different from that for the cohort sharing the first two phonemes (CV(*)) ($p < .01$), and for the cohort sharing the first phoneme (C(*)) ($p < .01$).

For the mean log frequency, the partial correlation coefficient for the cohort sharing the first three phonemes (CVC(*)) was significantly different from that for the cohort sharing the first two phonemes (CV(*)) ($p < .01$), and for the cohort sharing the first phoneme (C(*)) ($p < .05$).

For the maximum log frequency, the partial correlation coefficient for the cohort sharing the first three phonemes (CVC(*)) was significantly different from that for the cohort sharing the first two phonemes (CV(*)) ($p < .01$), and for the cohort sharing the first phoneme (C(*)) ($p < .01$).

Insert Figure 3 about here.

Discussion

Analysis 3 shows that the cohort sharing the first three phonemes (CVC(*)) is reliable in predicting word recognition performance in terms of the density, the mean log frequency, and the maximum log frequency. On the other hand, the cohorts sharing the first phoneme (C(*)) was not reliable in all cases. And the cohort sharing the first two phonemes (CV(*)) is reliable in terms of the density, but the partial correlation coefficient is much smaller than that found for the cohort sharing the first three phonemes.

These findings indicate that the cohort is reliable to the word recognition when all or a large part of it is included in the rime cognate. The present findings suggest that the cohort sharing the first two phonemes (CV*), which is usually used as the definition of the cohort, is in some degree reliable for the word recognition, because it contains the most distinctive part cohort (CVC*) which is included in the rime cognate. In other words, the reliability of the cohort in spoken word recognition derives from that of the rime cognate.

General Discussion

Analysis 1 showed that the rime cognate is a good predictor for lexical competition in spoken word recognition. It has much larger negative correlation than the neighborhood and the cohort. Analysis 2 and 3 showed that the rime cognate includes the parts of the neighborhood and the cohort which have greater negative correlation with recognition performance than other parts of them. In other words, the rime

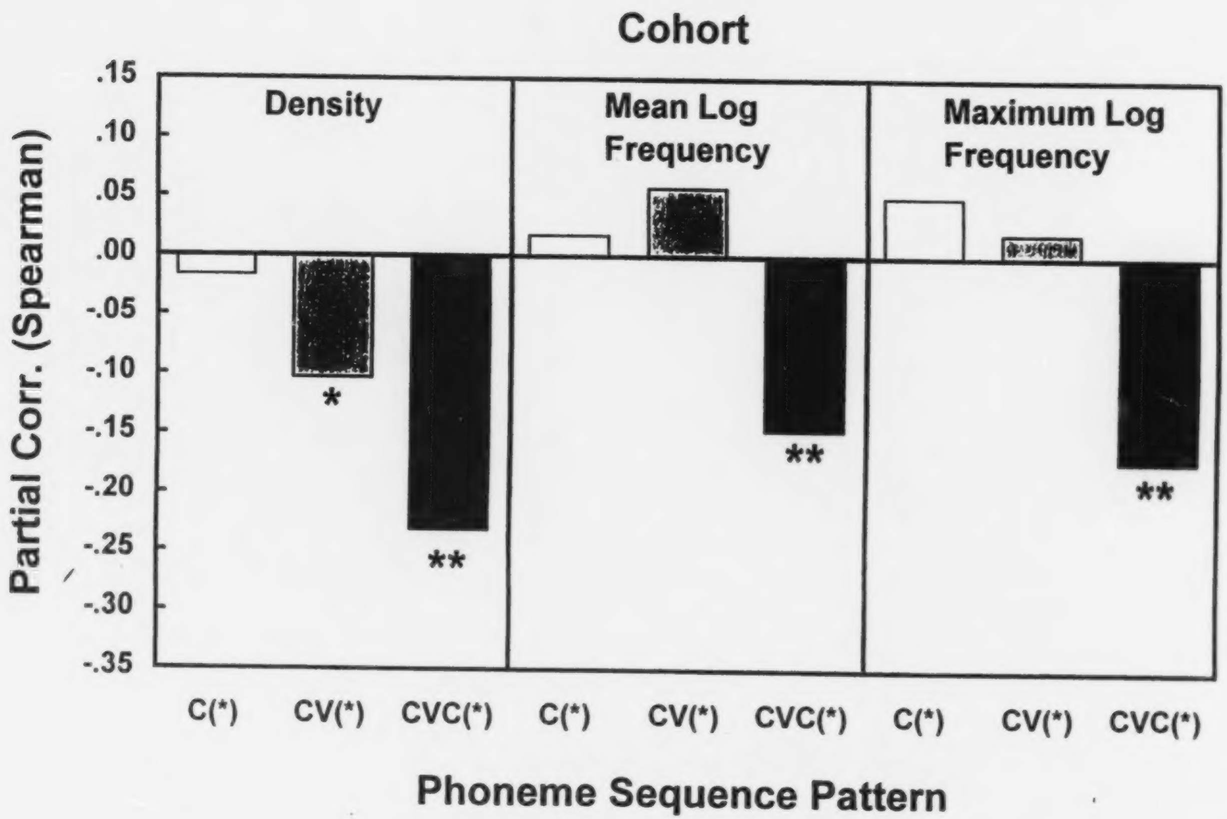


Figure 3. Spearman's rank order partial correlation coefficient between word recognition rate and density, mean log frequency, and maximum log frequency of cohort. Excluding factor for the partial correlation is log frequency of target word. "*" represents any length of phoneme sequence including null. "C" and "V" respectively represent a common consonant and vowel with the CVC target word. Single and double asterisks respectively represent difference from zero at 5% and 1% significance level.

cognate integrates the neighborhood and the cohort by discarding their unimportant parts. The integration provided by the rime cognate dissolves the discrepancy between the neighborhood and the cohort in the previous studies.

Strictly speaking, the reliability of the rime cognate is only shown for the CVC English words in this study. Thus, there are several unsolved questions. For example, is the rime cognate reliable for the words which do not have the CVC phoneme sequence pattern? In other words, monosyllabic words with more than two phonemes in the onset and/or the coda have not been investigated from the view point of the rime cognate. Multisyllabic words have not been investigated either.

However, there is some indirect support for that the rime cognate is a reliable word candidate set for non-CVC monosyllabic words. This comes from studies of phonological priming. Slowiaczek and Hamburger (1992) have shown that the prime word produces interference and increases reaction times of shadowing of the target word when there is an overlap of the initial three phonemes between the prime and the target word. In contrast, the prime word has no effects when the overlap is the first two phonemes, and it produces facilitative effects when the overlap is the initial phoneme.

Slowiaczek and Hamburger (1992) used monosyllables with four or five phonemes in length in their experiments. Phoneme sequence pattern of their stimuli and their proportion were CVCC (30%), CCVC (58%), CCVCC (11%), and CCCVC (1%) (L.M. Slowiaczek, personal communication, July 1, 1997).

Three phoneme overlap includes a part of the rime. That is, the vowel nucleus and the first phoneme of the coda for in CVCC pattern, and the vowel nucleus for CCVC and CCVCC pattern. The proportion of such stimuli was 99% in total. On other hand, two phoneme overlap only includes the vowel nucleus in CVCC pattern which is only 30% of the stimuli. And, one phoneme overlap did not include any part of the rime. Therefore, the possibility is larger for three phoneme overlap to include a part of the rime than one or two phoneme overlap.

Slowiaczek and Hamburger (Hamburger & Slowiaczek, 1996, Slowiaczek & Hamburger, 1992) argued that the interference in the three phoneme overlap reflects lexical competition, but the facilitative effect in the one phoneme overlap reflects prelexical processing. The interference in the three phoneme overlap can be interpreted as support for the proposal that the rime cognate is activated by a part of the rime and lexical competition among the rime cognate causes negative effects on phonological priming. However, because the rime cognate is activated by a part of the rime, not the whole rime, by this interpretation, it is a little bit different from the original definition of the rime cognate. Therefore, their results only indirectly suggest that the rime cognate is a reliable competitor set for the non-CVC monosyllabic words.

Some indirect support for reliability of the rime cognate on multisyllabic words comes from studies of word spotting. McQueen, Norris, and Cutler (1994) have shown that the multisyllabic word (domestic) which shares its beginning part with the carrier nonword (domes) competes with the monosyllabic target word (mess). Their results suggest that the monosyllabic target word and the multisyllabic word compete with each other because they are related with the rime. For example, "mess" and the second syllable of "domestic" share the rime of /es/. Hence, their results suggest that the rime cognate is activated at the second syllable in a multisyllabic word.

A study by Treiman, Fowler, Gross, Berch, and Weatherston (1995) provides additional support for the rime cognate in recognition of multisyllabic words. They conducted a series of experiments using word games in which a subject produced a new nonword by changing one or two phonemes in the noninitial syllable in a disyllable and trisyllable multisyllabic nonword. They found that subjects performed better in the games in which a changing unit corresponded to the onset and the rime than the games in which the changing unit does not correspond to the onset and the rime. Although their stimuli are nonwords, it is highly probable that the units of the rime and the onset also affect the recognition of a word with multisyllable structure. These findings suggest that the rime cognate may also be a reliable candidate set for multisyllabic words as well.

Further support for this proposal comes from the studies of the neighborhood of multisyllabic words. Cluff and Luce (1990) obtained recognition scores of two-syllable spondees words (e.g., "icecream", horseshoe). They manipulated the frequency, the neighborhood density, and the neighborhood frequency for each syllable in the word and found that the recognition scores of the two-syllable words are affected by these variables for each syllable. As shown in Analysis 2, a reliable part of neighborhood is included in the rime cognate. Therefore, if the neighborhood of each syllable affects the recognition of multisyllabic words, it is probable that the rime cognate of each syllable also affects it, too.

Taken together, the overall pattern suggests that spoken word recognition is affected by the rime cognate of non-CVC mono-/multi-syllabic words as well as the CVC monosyllabic words.

Another question that can be addressed here concerns the metrical stress and the rime cognate in multisyllabic words. Is the rime cognate activated by the rime of strong syllable, weak syllable, or both? The answer is probably that it is only activated by the rime of the strong syllable, because indirect evidence for the rime cognate has been obtained only in the strong syllable so far. For example, in McQueen, Norris, and Cutler (1994)'s study, competition was observed between "domestic" and "mess" which corresponds to the strong syllable of the "domestic." In Cluff and Luce's (1990) study, the spondees by definition consist of two strong syllables.

In addition, the strong syllable is more informative than the weak syllable because the number of vowel nuclei is larger in the strong syllable than in the weak syllable. These characteristics suggest that the weak syllable produces a larger rime cognate than the strong syllable. In other words, the weak syllable is not useful to reduce the number of word candidates in the rime cognate but the strong syllable does. Therefore, it is reasonable that the rime cognate is only activated by the rime of the strong syllable,

If the rime cognate is activated by the rime of the strong syllable with the rime matching strategy (RMS), it is important to consider the relationship to the metrical segmentation strategy (MSS) proposed by Cutler and Norris (1988). Although the MSS is used for prelexical processing, the MSS interacts with lexical processing (McQueen, Norris, & Cutler, 1994) The MSS is triggered by the beginning of a strong syllable. This is not consistent with the RMS which is triggered by the rime of the strong syllable. However, the MSS can be consistent to the RMS with a small modification. That is, the MSS can be modified to be triggered by the rime of the strong syllable. The justification for this modification is that there is not enough information in the consonant cluster of the onset to determine whether the syllable is strong or weak, and that the strong syllable is detected only after the vowel nucleus is processed. This modification permits the MSS be more realistic in the left-to-right processing and also be consistent to the RMS.

An example of this modification is shown as following. The recent version of SHORTLIST model (Norris, McQueen, & Culter, 1995) simulates the spoken word recognition incorporating the MSS by boosting the activation level of word units which have the strong syllable. In their simulation, the word candidates are boosted at the beginning of the strong syllable (i.e., at the onset of the strong syllable). However, if the modified MSS is applied, the word candidates should be boosted at the beginning of the rime of the strong syllable. In this case, boosted word candidates by the MSS will coincide to the rime cognate by the RMS.

Another question that needs to be addressed is the language dependency of the rime cognate. Is the rime cognate reliable in other languages? The answer might be negative. The rime cognate is probably reliable only in English (and some other language which use stress), because it may be activated by the rime of strong syllable. The strong syllable is a processing unit for segmentation in English (Cutler & Norris, 1988), but it is not for other languages. Different processing units may be in other languages. For example, it is a mora for Japanese (Otake, Hatano, Cutler, & Mehler, 1993; Cutler & Otake, 1994), and a syllable for French (Mehler, Dommergues, Frauenfelder, & Segui, 1981) and also for Spanish and Catalan (Sebastian-Galles, Dupoux, Segui, & Mehler, 1992). Although a specific processing unit for segmentation does not necessarily coincide to a lexical access unit, they are probably identical because of parsimony of processing. Therefore, the rime cognate which may depend on the strong syllable might not be reliable in other languages which do not use the strong syllable as a processing unit.

More direct negative evidence for the rime cognate has been obtained for Japanese. Watanabe (1996) measured the reaction times for discrimination between a target word and nontarget words which diverge from the target word at a certain phoneme in sequential matching. Using a regression equation for phoneme detection time (Amano, 1995), he found that the reaction times for the nontarget words have almost constant delay from detection time of the divergent phoneme at which the nontarget word can be discriminated from the target word in left-to-right matching procedure. His findings coincide with the original cohort model (Marslen-Wilson & Welsh, 1978) which claims that a word is processed in sequential matching and is perceived at the uniqueness point where the word diverges from all other word candidates. Although the divergent point in Watanabe's study does not correspond to the actual uniqueness point in a mental lexicon, his findings strongly suggest that the cohort, but not the rime cognate, is a reliable candidate set for recognition of Japanese spoken words.

Other studies implicitly assume that lexical access is universal among languages. However, it is highly probable that the lexical access is language dependent like segmentation of speech depends on language.

Based on the results and some considerations in this study, rime cognate model (RCM) is proposed for spoken word recognition as a set of working hypotheses. The RCM has following characteristics.

1. The rime of a given word is used as a cue for lexical access. Word candidates sharing the same phoneme sequence with the rime of a given word are activated when the rime is processed. This lexical matching procedure is called rime matching strategy (RMS). The set of word candidates which is activated by the RMS is called the rime cognate. The rime cognate is activated only by the rime in a strong syllable but not by the rime in a weak syllable. The rime ot given input word is matched to the phoneme sequence in every possible positions within a word in a mental lexicon.

2. The word candidates in the rime cognate compete against each other. Competition is realized by inhibitory information flow among word candidates. Competition continues until one candidate is highly activated and it exceeds a threshold for recognition.

3. Phonemes which are not in the rime of a strong syllable are used for increasing or decreasing the activation of word candidates. That is, facilitatory and inhibitory information are sent to word candidates from phonemes according to their consistency to words. This bottom-up information flow begins only after the rime cognate is developed by the RMS. Before that, no information flow exists from the phonemes to word candidates. The inhibitory information flow from phonemes to word candidates might be redundant if word candidates have inhibitory connections with each other. However, it may facilitate word recognition because irrelevant word candidates are deactivated sooner, which allows a relevant word candidate to be a winner more quickly.

4. A word candidate facilitates phoneme processing by changing phoneme's gain of activation. The word candidate makes the gain of the phoneme be greater in proportion to activation level of the word candidate when the phoneme is included in the word. These characteristics suggest that the facilitation on phoneme processing should be effective only after the rime cognate is developed, because the word candidate is not activated before it. And, the phoneme must have received some bottom-up information for being facilitated by the word, because the word only changed the gain of the phoneme but not the activation level itself.

5. Gain of activation of a word candidate is modulated by its word frequency and familiarity. A high-frequency/high-familiarity word has greater gain than a low-frequency/low-familiarity word, so that the former is more easily activated than the latter.

6. Syntactic and semantic information create a bias on the rime cognate by modifying the gain of word candidates. That is, syntactically and semantically correct word candidates are more easily activated than incorrect word candidates. This means that syntactic and semantic information indirectly reduces the number of word candidates after the lexical access is conducted by the RMS.

Some of the characteristics of the RCM may be easily implemented in the TRACE model (McClelland & Elman, 1986) and the SHORTLIST model (Norris, 1994; Norris, McQueen, & Cutler, 1995). For example, the RMS can be achieved by putting larger weights on the connections between the word units and the phoneme units which correspond to the rime. However, to precisely simulate the activation of a set for the rime cognate, they should have a gating mechanism which blocks information flow from phoneme units in the onset to word units until the rime is processed. This gating mechanism might be achieved by putting extremely large weights on the rime part and very small weights on the onset part. Then, word units which are slightly activated by the onset virtually would have null activation when word units are activated by the rime.

Other characteristics of the RCM might require a fundamental changes in the TRACE model and the SHORTLIST model. For example, the inhibitory connections between a phoneme unit and a word unit have not been implemented in either model. Effects of introducing inhibitory connection are unknown for the model's performance. However, the SHORTLIST model implicitly has such inhibitory connections by limiting the number of word candidates in word level.

The gain control might cause a fundamental change for these models, because they do not have such function. However, the gain control by top-down information has some merits which the activation control does not have in these models. That is, the top-down information affects a unit only if the unit receives some bottom-up information. This means bottom-up information dominates speech recognition. If top-down information affects a unit via adjustments in activation levels, the unit (i.e., a phoneme or a word) can be recognized without any bottom-up information when the top-down information is strong enough to fully activate the unit. This means that words can be recognized without speech input. The gain control can avoid such unrealistic situation.

The proposed characteristics of the RCM may have some superiority over the TRACE model and the SHORTLIST model. However, the set of hypotheses have not been confirmed. Further research is necessary to examine the predictions of the RCM.

In summary, a set of analyses were carried out using rime cognate as a candidate set for lexical processing in spoken word recognition. It is revealed that, at least for CVC English words, the rime cognate is more reliable lexical competitor set than the lexical neighborhood and the word-initial cohort, and that the rime cognate integrates the neighborhood and cohort by including their reliable part. It is suggested that the rime cognate is also a reliable candidate set for non-CVC words and multisyllabic words. But the rime cognate may be reliable to only spoken word recognition in English.

References

Amano, S. (1995). Time-course of phoneme/word perception in Japanese (*Technical Report of Hearing*, **H-95-50**). Tokyo: Acoustical Society of Japan, (in Japanese).

Bard, E. G. (1990). Competition, lateral inhibition, and frequency: Comments on the chapters of Frauenfelder and Peeters, Marslen-Wilson, and others. In G. T. M. Altmann (Ed.), *Cognitive models of speech processing: Psycholinguistic and computational perspectives* (pp. 185-210). Cambridge: MIT Press.

Bard, E. G., & Shillcock, R. C. (1993). Competitor effects during lexical access: Chasing Zipf's tail. In G. T. M. Altmann, & R. Shillcock (Eds.), *Cognitive models of speech processing: The second Sperlonga meeting* (pp. 235-275). Hillsdale: LEA.

Cairns, C. E., & Feinstein, M. H.(1982). Markedness and the theory of syllable structure. *Linguistic Inquiry*, **13**, 193-225.

Cluff, M. S., & Luce, P. A. (1990). Similarity neighborhoods of spoken two-syllable words: Retroactive effects on multiple activation. *Journal of Experimental Psychology: Human Perception and Performance*, **16**, 551-563.

Connine, C. M., Blasko, D. G., & Wang, J. (1994). Vertical similarity in spoken word recognition: Multiple lexical activation, individual differences, and the role of sentence context. *Perception & Psychophysics*, **56**, 624-636.

Cutler, A., & Norris, D. (1988). The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance*, **14**, 113-121.

Cutler, A., & Otake, T. (1994). Mora or Phoneme? Further evidence for language-specific listening. *Journal of Memory and Language*, **33**, 824-844.

Frauenfelder, U. H. (1990). Structure and computation in the mental lexicon. In H. Haken & M. Stadler (Eds.), *Synagetics of cognition* (pp. 406-414). Berlin: Springer-Verlag.

Frauenfelder, U. H. (1996). Computational models of spoken word recognition. In T. Dijkstra & K. de Smedt (Eds.), *Computational psycholinguistics* (pp. 114-138). London: Taylor & Francis.

Frauenfelder, U. H., Baayen, R. H., Hellwig, F. M., & Schreuder, R. (1993). Neighborhood density and frequency across languages and modalities. *Journal of Memory and Language*, **32**, 781-804.

Fudge, E. C. (1969). Syllables. *Journal of Linguistics*, **5**, 253-286.

Goldinger, S. D. (1989). Neighborhood density effects for high frequency words: Evidence for activation-based models of word recognition. *Research on Speech Perception, Progress Report* **15**, pp. 163-186. Bloomington, IN: Indiana University, Speech Research Laboratory.

Goldinger, S. D., Luce, P. A., & Pisoni, D. B. (1989). Priming lexical neighbors of spoken words: Effects of competition and inhibition. *Journal of Memory and Language*, **28**, 501-518.

Grosjean, F. (1980). Spoken word recognition processes and the gating paradigm. *Perception & Psychophysics*, **28**, 267-283.

Halle, M., & Vergnaud, J. (1980). Three dimensional phonology. *Journal of Linguistic Research*, **1**, 83-105.

Hamburger, M. B., & Slowiaczek, L. M. (1996). Phonological priming reflects lexical competition. *Psychonomic Bulletin and Review*, **3**, 520-525.

Kirtley, C., Bryant, P., MacLean, M., & Bradley, L. (1989). Rhyme, rime, and the onset of reading. *Journal of Experimental Child Psychology*, **48**, 224-245.

Kucera, H., & Francis, W. N. (1967). Computational analysis of present-day American English. Providence: Brown University Press.

Lenel, J. C., & Cantor, J. H. (1981). Rhyme recognition and phonemic perception in young children. *Journal of Psycholinguistic Research*, **10**, 57-67.

Luce, P. A. (1986). Neighborhoods of words in the mental lexicon. *Research on Speech Perception, Technical Report* **6**. Bloomington IN: Indiana University, Speech Research Laboratory.

MacKay, D. G. (1972). The structure of words and syllables: Evidence from errors in speech. *Cognitive Psychology*, **3**, 210-227.

Marslen-Wilson, W. (1984). Function and process in spoken word recognition—A tutorial review. In H. Bouma & D. G. Bouwhuis (Eds.), *Attention and performance* **X** (pp. 125-150). London: LEA.

Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word-recognition. *Cognition*, **25**, 71-102.

Marslen-Wilson, W. D. (1989). Access and integration: Projecting sound onto meaning. In W. Marslen-Wilson (Ed.), *Lexical representation and process* (pp. 3-24). Cambridge: MIT Press.

Marslen-Wilson, W. D. (1990). Activation, competition, and frequency in lexical access. In G. T. M. Altmann (Ed.), *Cognitive models of speech processing: Psycholinguistic and computational perspectives* (pp. 148-172). Cambridge: MIT Press.

Marslen-Wilson, W. D., Moss, H. E., & van Halen, S. (1996). Perceptual distance and competition in lexical access. *Journal of Experimental Psychology: Human Perception and Performance*, **22**, 1376-1392.

Marslen-Wilson, W. D., & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology*, **10**, 29-63.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, **18**, 1-86.

McQueen, J. M., Norris, D., & Cutler, A. (1994). Competition in spoken word recognition: Spotting words in other words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **20**, 621-638.

Mehler, J., Dommergues, J. Y., Frauenfelder, U., & Segui, J. (1981). The syllable's role in speech segmentation. *Journal of Verbal Learning and Verbal Behavior*, **20**, 298-305.

Norris, D. (1994). Shortlist: a connectionist model of continuous speech recognition. *Cognition*, **52**, 189-234.

Norris, D., McQueen, J.M., & Cutler, A. (1995). Competition and segmentation in spoken word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **21**, 1209-1228.

Nusbaum, H. C., Pisoni, D. B., & Davis, C. K. (1984). Sizing up the Hoosier mental lexicon: Measuring the familiarity of 20,000 words. *Research on Speech Perception, Progress Report 10*, pp. 357-376. Bloomington, IN: Indiana University, Speech Research Laboratory.

Otake, T., Hatano, G., Cutler, A., & Mehler, J. (1993). Mora or syllable? Speech segmentation in Japanese. *Journal of Memory and Language*, **32**, 258-278.

Pisoni, D. B., Nusbaum, H. C., Luce, P. A., & Slowiaczek, L. M. (1985). Speech perception, word recognition and the structure of the lexicon. *Speech Communication*, **4**, 75-95.

Sebastian-Galles, N., Dupoux, E., Segui, J., & Mehler, J. (1992). Contrasting syllabic effects in Catalan and Spanish. *Journal of Memory and Language*, **31**, 18-32.

Shillcock, R. (1990). Lexical hypotheses in continuous speech. In G. T. M. Altmann (Ed.), *Cognitive models of speech processing: Psycholinguistic and computational perspectives* (pp. 24-49). Cambridge: MIT Press.

Slowiaczek, L. M., & Hamburger, M. B. (1992). Prelexical facilitation and lexical interference in auditory word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **18**, 1239-1250.

Sommers, M. S. (1996). The structural organization of the mental lexicon and its contribution to age-related declines in spoken-word recognition. *Psychology and Aging*, **11**, 333-341.

Stemberger, J. P., & Treiman, R. (1986). The internal structure of word-initial consonant clusters. *Journal of Memory and Language*, **25**, 163-180.

Torretta, G. M. (1995). The "Easy-Hard" word multi-talker speech database: An initial report. *Research on Spoken Language Processing, Progress Report 20*, pp. 321-334. Bloomington, IN: Indiana University, Speech Research Laboratory.

Treiman, R. (1983). The structure of spoken syllables: Evidence from novel word games. *Cognition*, **15**, 49-74.

Treiman, R. (1985). Onsets and rimes as units of spoken syllables: Evidence from children. *Journal of Experimental Child Psychology*, **39**, 161-181.

Treiman, R. (1986). The division between onsets and rimes in English syllables. *Journal of Memory and Language*, **25**, 476-491.

Treiman, R., & Danis, C. (1988). Short-term memory errors for spoken syllables are affected by the linguistic structure of the syllables. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **14**, 145-152.

Treiman, R., Fowler, C. A., Gross, J., Berch, D., & Weatherston, S. (1995). Syllable structure or word structure? Evidence for onset and rime units with disyllable and trisyllable stimuli. *Journal of Memory and Language*, **34**, 132-155.

Treiman, R., & Zukowski, A. (1996). Children's sensitivity to syllables, onsets, rimes, and phonemes. *Journal of Experimental Child Psychology*, **61**, 193-215.

Vroomen, J., & de Gelder, B. (1995). Metrical segmentation and lexical inhibition in spoken word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, **21**, 98-108.

Watanabe, T. (1996). The reaction time for the discrimination of nontarget words. *The Journal of the Acoustical Society of Japan (E)*, **17**, 323-324.

Zwitserlood, P. (1989). The locus of the effects of sentential-semantic context in spoken-word processing. *Cognition*, **32**, 25-64.

Appendix

balm ban bead beak bean both bud bug bum bun cause chain chat
check cheer chief chore cod comb con cot curve dame death deep den
dirt does dog doom down dune fade faith fig fin firm five food
fool full gas gave girl give goat god gut hack hag hash hick
hid hoot hum hung hurl jack job join judge kin king kit knob
lace lad lame league learn leg lice live long lose love mace main
mall mat mid mitt moan moat mole mouth move mum neck noise pad
page pat path pawn peace pet pool pull pup put rat reach real
rhyme rim roof rough rout rum rut sane serve shall shape ship shop
sill size soak soil ·south suck tan teat teeth theme thick thing thought
toot vice voice vote wad wade wail was wash watch wed weed whore
wick wife work young

Some Computational Analyses of the PBK Test: Effects of Frequency and Lexical Density on Spoken Word Recognition[1]

Ted A. Meyer[2] and David B. Pisoni

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405

---

[1] This study was supported by NIH/NIDCD Grants DC00064, DC00111, and NIH Training Grant DC00012. Portions of the paper were presented at the 21st Annual Meeting of the Association for Research in Otolaryngology, St. Petersburg Beach, FL, February, 1998. We would like to thank Darla J. Sallee and Linette A. Caldwell for clerical assistance and Dr. Steven B. Chin for comments and suggestions on an earlier version of the manuscript.

[2] Department of Otolaryngology–Head & Neck Surgery, Indiana University School of Medicine, Indianapolis, IN.

Some Computational Analyses of the PBK Test: Effects of Frequency and Lexical Density on Spoken Word Recognition

Abstract. The Phonetically Balanced Kindergarten (PBK) Test (Haskins, 1949) has been used for almost 50 years to assess spoken word recognition performance in children with hearing impairments. The test originally consisted of four lists of 50 words, but only three of the lists (Lists 1, 3 & 4) were considered "equivalent" enough to be used clinically with children. Our goal was to determine if the lexical properties of the different PBK lists could explain any differences between the three "equivalent" lists and the fourth PBK list (List 2) that has not been used in clinical testing. Word frequency and lexical neighborhood frequency and density measures were obtained from a computerized database for all of the words on the four lists from the PBK Test as well as the words from a single PB-50 (Egan, 1948) word list. The words in the "easy" PBK list (List 2) were of higher frequency than the words in the three "equivalent" lists. Moreover, the lexical neighborhoods of the words on the "easy" list contained fewer phonetically similar words than the neighborhoods of the words on the other three "equivalent" lists. The present computational analyses show that both word frequency and lexical neighborhood density influence the probability of correct word recognition in open-set speech intelligibility tests. The results of this computational analysis of the PBK Test provide additional support for the proposal that spoken words are recognized "relationally" in the context of other phonetically similar words in the lexicon. Implications of using open-set word recognition tests with children with hearing impairments are discussed with regard to the specific vocabulary and information processing demands of the PBK Test.

Introduction

The PBK Test is an open-set test of spoken word recognition that has been widely used over the years in clinical audiology to measure speech perception skills in young children, especially deaf children with cochlear implants (CIs) and hearing aids (Carney et al., 1991; Fryauf-Bertschy, Tyler, Kelsay, & Gantz, 1992; Fryauf-Bertschy, Tyler, Kelsay, Gantz, & Woodworth, 1997; Gantz, Tyler, Tye-Murray, & Fryauf-Bertschy, 1994; Gantz, Tyler, Woodworth, Tye-Murray, & Fryauf-Bertschy 1994; Kirk, Osberger, & Pisoni, 1995; Miyamoto, Osberger, Robbins, Myres, & Kessler 1993; Miyamoto et al., 1994; Osberger et al., 1991; Pisoni, Svirsky, Kirk, & Miyamoto, submitted; Staller, Beiter, Brimacombe, Mecklenburg, & Arndt, 1991; Staller, Dowell, Beiter, & Brimacombe, 1991; Waltzman et al., 1994; Waltzman et al., 1995). The test was originally developed by Harriet Haskins in 1949 as part of her Master's thesis at Northwestern University to fill a need for an open set test of speech perception that could be used for children of kindergarten age. Although Haskins' thesis was never published, three of her original four word lists are routinely used clinically and have been employed extensively over the years to measure speech perception performance in young children with hearing losses.

As stated above, results from the PBK Test have come to play an extremely important role in studies of speech perception of children with CIs. Performance on the PBK Test is often used as the primary and "defining" criterion to identify exceptionally good users of CIs, the so-called "Star" performers (Pisoni et al., submitted). Children who display high levels of performance on the PBK Test also tend to perform well on other perceptual tests that are routinely administered as part of a standard assessment battery such as the Minimal Pairs Test (Robbins, Renshaw, Miyamoto, Osberger & Pope, 1988), Common Phrases Test (Osberger et al., 1991), Lexical Neighborhood Test (LNT) (Kirk et al., 1995), Peabody Picture Vocabulary Test (Dunn & Dunn, 1981), and the Reynell Developmental Language Scales (Reynell & Gruber, 1990). In addition, and perhaps even more interesting, recent analyses have shown that children who score exceptionally well on the PBK Test also display very good speech intelligibility as measured by transcription scores of naive listeners (Pisoni et al., submitted; Svirsky, 1996). This finding suggests that these particular deaf children have not only acquired the perceptual skills needed to recognize isolated words but have also developed the means to encode, represent and retrieve the sound patterns of the spoken words in memory. These children have acquired control over those aspects of expressive language that are needed to access motor patterns in speech production to produce intelligible speech. We believe these are important milestones in speech perception and language development, and they deserve more detailed examination in order to understand the basis for these achievements in deaf children with CIs.

Researchers and clinicians alike have routinely observed enormous amounts of variability among users of CIs on all of the standard assessment instruments (Carney et al., 1991; Fryauf-Bertschy et al., 1992; Fryauf-Bertschy et al., 1997; Gantz, Tyler, Woodworth et al., 1994; Staller, Dowell et al., 1991; Staller, Beiter et al., 1991). In the case of the PBK Test, the range of performance for children covers almost the full scale from zero to approximately 90 percent correct word recognition in some cases (Carney et al., 1991; Fryauf-Bertschy et al., 1992; Fryauf-Bertschy et al., 1997; Gantz, Tyler, Woodworth et al., 1994; Osberger et al., 1991; Waltzman et al., 1994). Upon closer inspection, the individual scores on the PBK Test tend to follow a bimodal distribution of performance: some of the children with CIs are able to perform well on the PBK open-set test whereas other children routinely score zero on the test (Lane, 1995). Why is there so much variability among CI users on the PBK Test? Why is the PBK Test apparently so difficult for some deaf children with CIs but much more manageable for other children? These are several of the questions we hope to answer in this paper.

In examining the task demands of the PBK Test, two major factors come to mind that are worth considering in some greater detail. First, we consider the specific vocabulary items used on this test. Another speech perception test for young deaf children was developed by Quick (1949) at the approximate time Haskins developed the PBK Test. Quick's closed-set test consisted of two lists of 25 monosyllabic words, with each word being accompanied by two very similar-sounding words. All of the English speech sounds were represented in the two lists, but the two lists did not maintain the phonetic balance of the Phonetically Balanced (PB) lists that were developed earlier in the Psycho-Acoustic Laboratory at Harvard University (Egan, 1948). Haskins originally developed the PBK lists from words taken from the well-known PB-50 lists for the adult listeners, but also used words that were in the 2500 words of highest frequency of words spoken by preschool children (The International Kindergarten Union, 1928). It has been suggested recently by Kirk et al. (1995) that children with limited vocabulary skills, such as very young children or children with CIs, may score poorly on or be unable to do the PBK Test at all because the specific words used on the test are simply too difficult and are not within the vocabulary of these children. Is this explanation of the difficulty of the PBK Test correct? How can this account be assessed?

At the present time, there are no vocabulary norms or lexical databases for deaf children, so it is difficult to compare the words on the PBK Test to words from the vocabularies of deaf children. Kirk et al. (1995) attempted a computational analysis of the words on the PBK Test using a lexical database constructed by Logan (1992). Logan's database contained samples of the vocabulary from normal-hearing children that was available in the CHILDES (Child Language Data Exchange System) database (MacWhinney & Snow, 1985). Kirk et al., however, found that a large number (approximately 69%) of the words on the PBK Test were, in fact, not within Logan's database. Thus, it is possible that the children

BEST COPY AVAILABLE

with CIs cannot recognize the words on the PBK Test simply because they are not familiar with the specific items used in this test?

In addition to the specific vocabulary items used on the PBK Test, the lexical properties of the words themselves may also play an important role in recognition. Some words may be hard to perceive because the children simply do not know them. Alternatively, some words may be hard to perceive because they are phonetically confusable with many other similar-sounding words (see Luce & Pisoni, in press). And, some words may be hard to perceive because deaf children have difficulty making fine phonetic discriminations that are needed for identification of these particular sound patterns. In their computational analyses of the vocabulary of the PBK Test, Kirk et al. (1995) found that a large number of words on the PBK Test were perceptually difficult; that is, the words came from regions of the lexicon where there were many other phonetically-similar words that were higher in frequency than the target word. These observations suggest that discriminability and competition among lexical candidates that are phonetically similar to each other may also affect performance on the PBK test and may be another important factor that makes the words on this test difficult for deaf children with CIs (see Luce & Pisoni, in press). Thus, the PBK Test may be a very difficult test for young deaf children not only because many of the words are unfamiliar and therefore not part of their lexicon but also because many of the words are extremely difficult to perceive in isolation where the only available context is the presence of other phonetically-similar words in the child's language.

In addition to the specific vocabulary items used on the PBK Test, it is also possible that the specific task demands of an open-set word recognition test also affect performance, particularly in young deaf children who may not be able to make fine phonetic discriminations among different sound patterns and encode these into memory. An open-set word recognition test like the PBK Test has no external context or response constraints that a listener can rely on other than the knowledge of sound patterns and regularity of words in their lexicon. Recognizing a word in isolation therefore requires that the listener encode the sound pattern into working memory, and then access a motor program for the word from information stored in the lexicon in order to produce an utterance on demand in a repetition task. Children may therefore have difficulty with an open-set test not only because the vocabulary is unfamiliar but because they are not able to encode, represent or access the sound pattern of these novel items from memory in what is essentially an imitation task. Either or both of these alternatives are possible accounts of why the PBK Test is difficult for young deaf children. Because of the importance placed on open-set word recognition tests as a criterion for evaluating performance with CIs, we believe that it is critical at this time to gain an understanding of the reasons for the generally poor performance observed on the PBK test with hearing-impaired children and to analyze the specific task demands of this test.

After reading the original version of Harriet Haskins' thesis, we discovered several interesting findings that we wish to report here both for historical reasons, because her thesis was never published in a scientific journal and therefore has had only a very limited distribution within the hearing and speech science community, and for scientific reasons, because her original results obtained almost 50 years ago are still informative and bear on several current theoretical issues concerning speech perception and spoken word recognition in young children. In particular, based on our analyses reported below, we believe it is possible to provide a principled explanation of why the PBK Test is a difficult test for some children with CIs. Several new computational analyses of the specific words used on the PBK Test were conducted, and these results provide some new insights into the primary factors that influence word recognition in open-set tests. In the sections below, we first describe how the PBK word lists were constructed and summarize Haskins's findings. Then we report the results of our computational analyses of the word lists. Finally, we propose an explanation of the pattern of her results and discuss the implications of these new findings for assessing speech perception and spoken word recognition in deaf children with CIs.

Although not widely known, Haskins actually constructed *four* separate lists of PBK words and then collected speech intelligibility data on these materials which were included in her thesis. However, only three of these original lists are currently being used by audiologists in the clinic and in research laboratories. Apparently, the fourth PBK list was never used because, according to Haskins, her speech intelligibility scores with adult listeners with normal hearing showed that the words on this particular list (PBK List 2) were "more audible" that the words used on the other three lists. In order to identify the basis for these differences, we have carried out a series of "computational" analyses of the words used in all four of Haskins' word lists and we report the results below. We believe these new findings on the words used in the PBK Test are important and should be of interest to researchers and clinicians who use speech intelligibility tests to measure and assess changes in speech perception and word recognition skills in young children, particularly young deaf children with hearing aids or CIs.

As mentioned previously, Haskins collected speech intelligibility data with these four word lists using normal-hearing adults as subjects. To the best of our knowledge, no speech intelligibility data were ever collected from young children with normal hearing, the target population that the materials were originally developed for at the time, nor have any speech intelligibility data been published from children with normal hearing using these lists. Because of the important role that the PBK Test has played in recent years in assessments of speech perception and word recognition skills in young children, we critically examine both issues below.

In the first section of this paper, we report new computational analyses on all four of Haskins' original word lists. The results of these analyses permit us to offer a theoretically-motivated explanation for the differences in speech intelligibility she observed between the lists of words with adult listeners. Not only can we now account for differences among the four original lists, but these new computational analyses of Haskins's data also provide additional support for a recent model of spoken word recognition – the Neighborhood Activation Model (NAM) (Luce, 1986; Luce & Pisoni, in press; Luce, Pisoni & Goldinger, 1990). More importantly, the results of these analyses provide new evidence for the proposal that spoken words are recognized "relationally" in the context of other phonetically similar words in the listener's lexicon. Using the Neighborhood Activation Model of spoken word recognition as our theoretical framework, we are able to offer an explanation for why some words are easy to recognize and why other words are more difficult to recognize in an open-set test format using a few simple principles that characterize the recognition process for spoken words (see Luce & Pisoni, in press). Finally, we offer some comments about the implications of these findings for assessment of speech perception in deaf children.

Methods

Word Lists

Haskins recorded a single randomization of the four lists of the PBK Test. She then presented the four PBK lists and one of the PB-50 lists (List 13) at sequentially higher signal levels to adult listeners with normal hearing. The highest level was approximately 27 dB SL (0 dB SL was determined by having the subjects set an attenuator to a level at which they felt they were obtaining approximately 50% of a sample of continuous discourse). Listeners were asked to identify the words using an open-set format. Haskins found that List 2 was "easier" than the other lists at attenuation levels of 50 (12 dB SL) and 60 (2 dB SL) dB, but at the highest level (27 dB SL) the scores from all five lists were essentially equivalent, due to a ceiling effect. Haskins stated that she felt that List 2 was "easier" than the other lists at all levels of presentation, although a significant difference between the lists emerged only at 50 and 60 dB attenuation. She also stated that an "item analysis" of the individual words in PBK List 2 might yield additional information on ways to improve the list, or to make it more like the other lists, but she did not indicate what kind of item analysis to do or what perceptual dimensions of spoken words might be relevant to speech intelligibility performance or to the differences in performance she observed.

Lexical Neighborhood Database

The four PBK lists of Haskins (1949) and List 13 of the PB-50 words (Egan, 1948) were analyzed using several techniques to compute similarity spaces for spoken words (see Pisoni, Nusbaum, Luce, & Slowiaczek, 1985). A computational analysis of the specific words used on all five lists was carried out using a computerized version of a 20,000-word Webster's Pocket Dictionary (Pisoni et al., 1985). The word lists are given in Table 1. From the pocket dictionary, we obtained word frequencies (Kucera & Francis, 1967), as well as lexical neighborhood frequencies and densities. Word frequency refers to the frequency counts from the Kucera and Francis (1967) norms. A "lexical neighborhood" of a word is defined as all words in the pocket dictionary that differ from the stimulus word by a single phoneme *substitution, addition,* or *deletion* (Greenberg & Jenkins, 1963). For example, if the stimulus is "pit", an example of a neighbor by *substitution* is "bit", by *deletion* "it", and by *addition* "spit". Neighborhood density refers to the number of words in the lexical neighborhood, and neighborhood frequency refers to the frequency counts (Kucera and Francis, 1967) of the words in the lexical neighborhood. The mean word frequency was calculated for all the words in each list. Because a few of the words had very high frequencies (e.g., "and" in List 2), the median word frequency was also calculated for each list. The median as a measure of central tendency tends to be less dependent upon outliers in asymmetrical or skewed distributions. Mean and median lexical neighborhood frequencies and densities were also calculated for the words in each individual list. Two additional measures were calculated for each word, the ratio of the word frequency to the neighborhood frequency and the ratio of the word frequency to the neighborhood density. These "second-order statistics" provide measures of the *relation* of a specific word to its neighbors and are used to quantify the amount of lexical competition among phonetically similar words in lexical memory.

Results

The results of the computational analyses for the four PBK lists (1-4) and the single PB-50 list (13) are shown in Tables 2 and 3. Haskins (1949) found that the words on List 2 were easier to recognize at several S/N than the words on the other lists, and she concluded that the items on List 2 were "more audible" than the words on the other three PBK lists at two intensity levels. Haskins obtained percent correct scores from 22 normal-hearing adults for the PBK words. The word lists are presented in Table 1, and the results from these intelligibility tests are shown in Table 2 (adapted from Haskins, 1949, Table VI). According to Haskins, at the middle intensity tested (2.2 dB SL), List 2 is significantly easier than Lists 1, 3, and 13. At the next higher intensity tested (12.2 dB SL), List 2 is significantly easier than all the other lists. At the highest intensity tested (27.2 dB SL), List 2 is significantly easier than only List 13, but at this level, performance is nearly perfect, and any differences between lists are confounded by ceiling effects.

Table 1.

Words on PBK Lists 1-4 and PB-50 List 13.

	PBK-1	PBK-2	PBK-3	PBK-4	PB-13
1	please	this	laugh	tire	bat
2	great	ma	falls	seed	beau
3	sled	pick	paste	purse	change
4	pants	glove	plow	quick	climb
5	rat	gun	page	room	corn
6	bad	forth	weed	bug	curb
7	pinch	trade	gray	that	deaf
8	such	each	park	sell	dog
9	bus	ask	wait	low	elk
10	need	wake	fat	rich	elm
11	ways	calf	ax	those	few
12	five	rope	cage	ache	fill
13	mouth	night	knife	black	fold
14	rag	chew	turn	else	for
15	put	guess	grab	nest	gem
16	fed	wave	rose	jay	grape
17	fold	cloud	lip	raw	grave
18	hunt	good	bee	true	hack
19	no	barn	bet	had	hate
20	box	left	his	cost	hook
21	are	shoe	sing	vase	jig
22	teach	flag	all	press	made
23	slice	rode	bless	fit	mood
24	is	hook	suit	bounce	mop
25	tree	front	splash	wide	moth
26	smile	toe	path	most	muff
27	bath	south	feed	thick	mush
28	slip	rest	next	if	my
29	ride	tongue	wreck	them	nag
30	end	best	waste	sheep	nice
31	pink	reach	crab	air	nip
32	thank	slide	peg	set	ought
33	take	food	freeze	dad	owe
34	cart	new	race	ship	patch
35	scab	ball	bud	case	pelt
36	lay	three	darn	you	plead
37	class	closed	fair	may	price
38	me	kept	sack	choose	pug
39	dish	off	got	white	scuff
40	neck	sick	as	frog	side
41	beef	thread	grew	bush	sled
42	few	day	knee	clown	smash
43	use	feel	fresh	cab	smooth
44	did	wood	tray	hurt	soap
45	hit	pig	cat	pass	stead
46	pond	crack	on	grade	taint
47	hot	dime	camp	blind	tap
48	own	wash	find	drop	thin
49	bead	and	yes	leave	tip
50	shop	look	loud	nuts	wean

Haskins' actual data (Percent Correct vs Sensation Level) are replotted in Figure 1 (adapted from Haskins, 1949, Figure 7). The data points as well as the best-fitting sigmoidal functions are plotted in this figure. The best-fitting curves are described as follows:

$$P(C)=\frac{a}{1+e^{-\frac{1}{b}(level-level_0)}}$$

where P(C) is the percentage of correct responses, a is the maximum possible P(C), $level_0$ is the level at the midpoint of the function, and b is the slope of the curve at the midpoint. The five curves in Figure 1 correlate with the individual data points nearly perfectly, ($r > +.99$).

Table 2.

Mean percent correct responses and standard deviations for PBK Lists 1-4 and PB-50 List 13. Values averaged across 22 adult listeners with normal hearing (adapted from Haskins, 1949, Table VI).

		PBK List 1		**PBK List 2**		**PBK List 3**		**PBK List 4**		**PB-50 List 13**	
Attenuation Level	Sensation Level	Mean	S.D.	Mean	S.D.	Mean	S.D	Mean	S.D	Mean	S.D.
70	-7.8	3.4	3.8	7.6	8.9	6.7	6.3	6.8	5.7	5.7	5.6
65	-2.8	16.3	10.3	24.2	10.8	16.8	8.9	19.5	9.1	17.5	8.2
60	2.2	33.8	10.6	44.1	11.9	36.6	9.5	39.3	12.5	33.4	8.2
50	12.2	69.9	8.8	82.2	11.2	71.5	11.2	73.2	9.0	68.2	7.8
35	27.2	98.1	3.1	98.4	1.7	97.1	2.8	97.7	2.4	95.7	1.9

Table 3.

Lexical neighborhood measures for PBK Lists 1-4 and PB-50 List 13.

	PBK List 1	**PBK List 2**	**PBK List 3**	**PBK List 4**	**PB-50 List 13**
Mean (Word Frequency)	558.6	954.4	689.1	612.7	295.6
Mean (Neighborhood Frequency)	294.6	261.4	231.5	272.5	196.9
Mean (Neighborhood Density)	17.0	16.4	17.5	17.7	17.3
Mean (Word Frequency/Neighborhood Frequency)	4.4	8.0	2.0	3.3	1.9
Mean (Word Frequency/Neighborhood Density)	35.3	94.1	42.5	41.3	13.0
Median (Word Frequency)	62.0	110.0	46.0	75.0	13.0
Median (Word Frequency/Neighborhood Frequency)	0.9	1.4	0.9	0.9	0.2
Median (Word Frequency/Neighborhood Density)	3.9	5.3	2.7	5.0	0.7

In examining the curves in Figure 1, one can see that performance on the PBK Test is clearly related to the presentation level of the stimulus. The slopes of the articulation functions between 20-80 percent correct response are approximately 4%/dB. At the ends of the functions (0-20%, 80-100%), performance changes little with a change in the level of presentation (approximately 1%/dB). Thus, for listeners with hearing losses, increasing the presentation level of the stimulus words by 5 dB could amount to as much as a 20% increase in performance, or very little increase in performance depending upon how much useable hearing the subject has and what region of the curve the subject is working in.

Insert Figure 1 about here

Haskins also examined test-retest reliability for her word lists at a single presentation level (12.2 dB SL). She found a large amount of improvement between the first and second repetitions of the test with the average score for List 2 increasing approximately 5.5%, and the average score for the remaining lists increasing approximately 8.8%. Significant improvement was seen for Lists 1 and 4 ($p < .05$), Lists 3 and 13 ($p < .01$), but not for List 2 ($p > .05$).

The results of the computational analysis of the lexical properties of the words on the PBK lists are displayed in Table 3. Here, we show both the mean and median word frequency, neighborhood frequency, and neighborhood density for the test items on the five word lists. In examining the mean word frequency, the reader will notice a very large difference in the mean frequency between List 2 (954.4) compared to the mean frequencies of the other three lists (295.6 - 689.1). Upon closer inspection, List 2 contains the word "and" which is one of the six most common words in the English language. The frequency of "and" is very high and skews the frequency distribution of List 2 a great deal. However, if we examine the median word frequency for the different lists, we still find that List 2 has a higher median frequency (110) than the other three lists (13 - 75). Thus, the words in List 2 do occur more frequently in the language than the words in the other lists, which may account, in part, for why List 2 was "more audible" than the other PBK lists.

Little difference exists in the mean neighborhood frequencies or densities between List 2 and the other test lists. The mean neighborhood frequency of List 2 (261.4) is within the range of mean neighborhood frequencies for the other three lists (196.9 - 294.6). The mean neighborhood density of List 2 (16.4 words) is slightly less than the mean neighborhood densities of the other lists (17.0 - 17.7 words). However, these differences are not very large. Thus, it appears that the mean neighborhood frequency and density of the words for the different PBK lists do not provide us much useful information about the reasons for the differences in audibility observed between the lists. However, these three measures (word frequency, neighborhood frequency, and neighborhood density) are absolute values for the words themselves without regard to context and the effects of other phonetically similar words. Put another way, these are "first-order" computational measures that do not take into account the relational properties of words to other phonetically similar words in the lexicon.

The next step in the computational analysis of the lexical properties of the PBK words was to generate ratios of the individual word frequencies to the neighborhood frequencies and densities of the individual words. These measures are "second-order" statistics that capture the relational properties of words to their lexical neighborhoods. Both mean and median values are listed in Table 3, but because of the effect of very high-frequency words (e.g., "and"), we will just examine the median values. Ratios were generated by dividing the frequency of the individual words by the neighborhood frequency of the

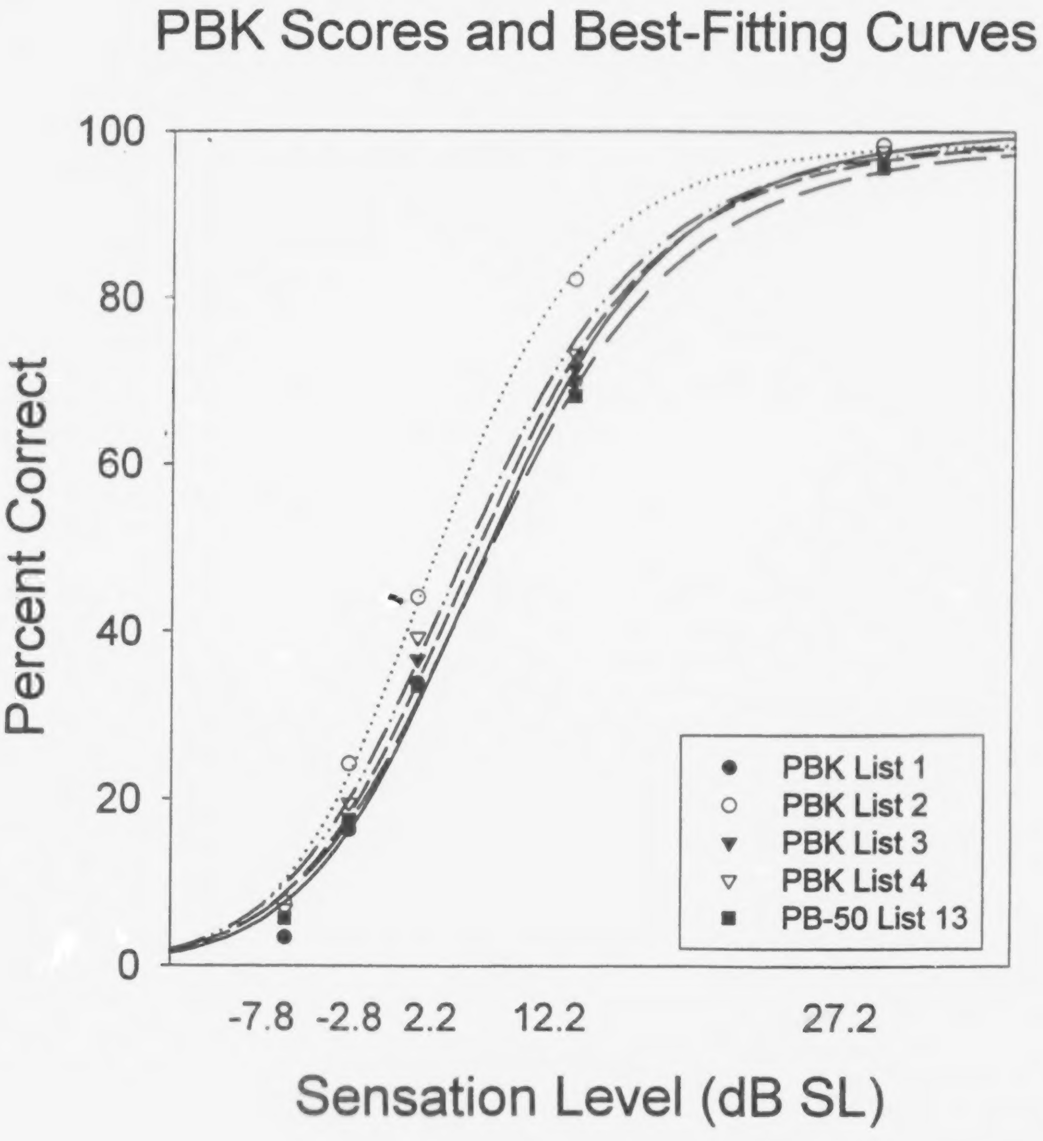


Figure 1. Percentage of correct responses vs Sensation Level for PBK Lists 1-4 and PB-50 List 13. Best-fitting curves are 3-parameter sigmoidal functions described in Equation 1. Adapted from Haskins (1949) Figure 7.

356

individual words. The median of these ratios for List 2 was approximately 1.4 while the median ratios for the three other PBK lists (0.2 - 0.9) were much smaller than the median ratio for List 2. This finding demonstrates that the words in List 2 have a higher frequency of occurrence than the other words in their lexical neighborhoods whereas the words in the remaining lists have a lower frequency of occurrence than the words in their lexical neighborhoods. Put another way, this analysis demonstrate that there is less lexical competition among phonetically similar words on PBK List 2 than the words on the other three PBK lists.

A second statistic was generated by dividing the frequency of the individual words by the neighborhood density of the individual words. The results of this analysis showed that the median ratio for List 2 (5.3) was larger than the median ratios for the other three PBK lists (0.7 - 5.0). As reported earlier, the median neighborhood densities for the different PBK lists were quite similar (range 16.4-17.7 words), however, this statistic provides further evidence that there is less lexical competition among phonetically similar words on PBK List 2 than the words on the other three PBK lists. Thus, words on List 2 should be less confusable than words on the other three lists. In short, these computational analyses of the similarity spaces or lexical neighborhoods of the words demonstrate that the words in PBK List 2 are inherently more distinctive and discriminable than the words on the three other PBK lists. If we assume that in open-set tests spoken words are recognized relationally in the context of other phonetically similar words, these computational analyses would predict that the words on PBK List 2 should be more acoustically distinctive and therefore more easily recognizable than the words on any of the other three lists. This is precisely the result that Haskins found in her thesis in 1949 almost 50 years ago.

Discussion

The results from the present analysis of the specific words used on the PBK Test give further support to NAM (Luce, 1986; Luce & Pisoni, in press; Luce, Pisoni, & Goldinger, 1990). NAM is a model of speech perception and spoken word recognition whose underlying hypothesis is that words are recognized relationally in the context of other phonetically similar words. This model attempts to account for the effects of lexical activation and competition on spoken word recognition. In particular, it is assumed that a sound pattern activates multiple lexical items in memory and this pattern of activation affects lexical discrimination and subsequent recognition performance (Luce & Pisoni, in press).

NAM predicts that when a spoken word is perceived, it activates a set of representations of similar sounding words in memory. This set of words is called a "lexical neighborhood." Once the neighborhood of the stimulus is activated, the word recognition system must then discriminate among the activated neighbors or candidates and decide which of the neighbors best matches the stimulus. The lexical discrimination and decision process is influenced by three factors: (1) the frequency of the target word; (2) the number of phonetically similar words (neighbors) activated in memory by the stimulus pattern; and (3) the frequencies of occurrence of the neighbors in the similarity neighborhood. Because of increased competition among activated items in memory, stimuli from densely populated lexical neighborhoods are more slowly and less accurately recognized than stimuli from sparsely populated neighborhoods. In addition, the presence of high frequency neighbors produces increased competition among words that are similar to the target word, thereby also reducing recognition performance.

According to NAM, identification performance can be predicted by the following rule (from Luce, 1986):

$$p(ID)=\frac{p(Stim)*Freq_s}{p(Stim)*Freq_s+\sum_{j=1}^{n}p(Neighbor_j)*Freq_j}$$

where p(ID) is the probability of correction identification, $p(Stim)*Freq_s$ is the frequency-weighted probability of identifying the stimulus word based on acoustic-phonetic information, and $p(Neighbor_j)*Freq_j$ is the frequency-weighted probability of identifying a neighbor. According to this rule, the model predicts that stimulus words with few low-frequency neighbors will be identified most accurately (because of the relatively small denominator in Equation 2, which indexes the degree of lexical competition among neighbors activated in memory). The model also predicts that stimulus words with many high-frequency neighbors will be identified least accurately (relatively large denominator in Equation 2).

The present analysis of the items on the PBK Test revealed that the median values of word frequency for the words in List 2 are considerably higher than the median values of word frequency for the words in the other lists. According to Pollack, Rubenstein and Decker (1959), listeners are more likely to correctly identify a high-frequency word than a low-frequency word at unfavorable S/N ratios when the message set is unknown. Thus, one would expect, and NAM predicts that on the average, at unfavorable S/N ratios, listeners would be more likely to correctly identify a word from List 2 than a word on the other lists. Moreover, scores with List 2 should be better than scores for the other lists at points along portions of the psychometric functions seen in Figure 1. At the end of the functions where the sensation levels of the stimuli are very low, the effect of the frequency of the stimulus word on the identification of that word is not very important, because there is little useful information in the signal at a presentation level at which some of the words are just becoming intelligible (very low S/N ratio). At the other end of the psychometric functions, the presentation level of the stimuli is high enough (high S/N ratio) so that all the words are easily intelligible, and word frequency again has little effect on identification.

The results of our analyses also revealed that the ratio of the frequency of the stimulus to the frequency of the neighborhood of the stimulus was higher for List 2 than the other lists. If the frequency of the stimulus is relatively great compared to the frequency of the neighborhood of that stimulus, then the numerator of Equation 2 is relatively large compared to the denominator, and the probability of correctly identifying the stimulus should high. Again, this effect is more important at intermediate S/N ratios where the stimulus is neither too easy nor too difficult to identify. The ratio of the frequency of the stimulus to the density of the neighborhood was also more favorable for the words on List 2 than the words on the other three PBK lists. If the density, or number of items in a lexical neighborhood is low in relation to the frequency of the stimulus, then the denominator in Equation 2 is again relatively small when compared to the numerator, and the probability of a listener correctly identifying the stimulus should be high. Thus, we find that both absolute and relative lexical neighborhood measures obtained in the present study lend further support to NAM.

The findings reported here are of both clinical and theoretical interest. A more practical question arises from the results -- are the lexical neighborhood differences observed here important for the target group the test was designed for and under the conditions the test is normally given? First, as was noted earlier, the PBK Test was constructed to be a test of word identification for children, and especially children with hearing losses. However, Haskins measured the "equivalence" of the PBK lists using adult subjects with normal hearing, not children with normal hearing or hearing loss. Nevertheless, the PBK Test has proven to be a very useful test for children, and it has probably been one of the most important measures of open-set word recognition for children with CIs (Carney et al., 1991; Fryauf-Bertschy et al., 1992; Fryauf-Bertschy et al., 1997; Gantz, Tyler, Tye-Murray et al., 1994; Gantz, Tyler, Woodworth et al., 1994; Kirk et al., 1995; Miyamoto et al., 1993; Miyamoto et al., 1994; Osberger et al., 1991; Pisoni et al., submitted; Staller, Beiter et al., 1991; Staller, Dowell et al., 1991; Waltzman et al., 1994; Waltzman et al., 1995). Second, the PBK Test is usually given by an audiologist through either monitored live voice or a tape-recorded version of the test at approximately 70 dB SPL. The most intense level Haskins presented to her subjects was approximately 27 dB above the average spondee thresholds for her group of subjects. If we assume that an average Speech-Reception Threshold (SRT) for spondaic words is approximately 20 dB SPL (ANSI S3.6, 1989; Young, Dudley, & Gunter, 1982) for a group of young adults with normal hearing, then Haskins presented the PBK lists to her subjects at approximately 47 dB SPL. We would expect performance to be perfect or nearly perfect for normal-hearing adults at a presentation level of 70 dB SPL, regardless of the PBK list tested, and that no differences between the lists would be apparent at this level.

For children with normal hearing, we would also expect nearly perfect performance at 70 dB SPL assuming that the children were able to perform the task and actually knew the words being tested. For children with hearing losses, their performance should be related to the level at which the word lists are presented. If the child has useable hearing and is able to understand speech well, and if the PBK words are presented at a high enough level, the child should perform well (nearly 100%) on the PBK Test (and performance should be nearly identical for the four different PBK lists). If the PBK lists cannot be tested at an appropriate level, or the child does not have useable hearing, then we would expect performance on the PBK Test to be low, and we might expect to see differences emerge between scores on PBK List 2 and the other PBK lists based on the lexical differences between the lists reported above.

The words chosen for the PBK Test were selected to be within the vocabulary of a child of based on norms for preschool children (The International Kindergarten Union, 1928). As stated earlier, Kirk et al. (1995) found that a large number of words on the PBK Test were not in a database containing words in the vocabulary of young children (Logan, 1992). In a more recent study, Kluck, Pisoni and Kirk (in preparation) obtained speech perception scores from 22 three-year-olds and 8 four-year-olds using recorded versions of the PBK Test and the LNT. The words were presented to the children at approximately 70 dB SPL in an auditory-only format. The four-year-old children performed at near-ceiling levels for both tests with performance on the PBK Test just slightly lower than performance on the LNT. Performance for the three-year-old children was slightly lower than performance for the four-year-old children on both tests. These results from the children with normal hearing provide a benchmark for children with impaired hearing and CIs. Clearly, three- and four-year-old children with normal hearing are able to imitate and repeat all the test items on both lists without any difficulty.

Kluck et al. (in preparation) addressed the issue of word familiarity in the different lists by asking the parents of the children whether or not their child was familiar with the word, using a seven-point scale. Although Kirk et al. (1995) found that a large number of words on the PBK Test were not in a corpus of words based on the vocabularies of young children (Logan, 1992), Kluck et al. found that all the words on both the PBK Test and the LNT were reported to be very familiar to the 4-year-old children, with the mean familiarity rating of the words from the PBK Test rated 6.58 (out of 7) and the words from the LNT rated 6.87 (out of 7). The words were slightly less familiar to the 3-year-old children (mean familiarity of PBK words - 5.95, mean familiarity of LNT words - 6.74 out of 7). Thus, it appears that the words on the PBK Test are less familiar to the three-year-olds than to the four-year-olds, but most of the words on the PBK Test are familiar to preschool-aged children.

Although the PBK Test may have certain shortcomings, it is certainly a difficult and probably the most important test of open-set word recognition for deaf children both with and without CIs. The tasks involved in the correct repetition of a stimulus in an open-set test of word recognition place a number of information processing demands on deaf children. The listener must search and retrieve words from lexical memory by encoding phonemic differences based on information present in the speech signal without the aid of any external context or retrieval cues and then discriminate and select a pattern from a large number of equivalence classes in memory. The listener must then use this information to form an articulatory plan or a motor program in order to produce a verbal or motor response to the stimulus. The various perceptual and cognitive operations needed for spoken word recognition require access to a variety of memory codes and neural representations of speech and spoken language at different levels of analysis. The speed and efficiency of these information-processing operations, particularly as they might be employed in tasks requiring transformation and mapping from perception to production, will depend to a large extent on having representations of words in memory and organizing these representations systematically in a lexicon that can be accessed efficiently to provide different sources of information about the words in the language.

We know that pediatric CI users recognize words relationally in accordance with the predictions of NAM. This was demonstrated by Kirk et al. (1995) using the LNT, which was designed especially to examine this question. It is possible that the lexical neighborhoods of children with CIs are fundamentally different from the lexical neighborhoods of children with normal hearing. We predict that because of poorer discrimination abilities, the neighborhoods of the CI users will be broader (i.e., contain more words) than the neighborhoods of children with normal hearing. For example, let us assume that a CI user has difficulty discriminate the place of articulation of a consonant. If the stimulus word presented to the listener is "pan", that listener would have difficulty discriminating the stimulus from words that vary from the stimulus word in the place of articulation of the initial consonant (e.g., "tan" and "can"). For a listener with normal hearing, the lexical neighborhood of the word "pan" includes the words "tan" and "can" as they vary from the stimulus by a single phoneme substitution. For the CI user who has difficulty discriminating the place of articulation of consonants, the lexical neighborhood of the stimulus "pan" might also contain words in the lexical neighborhoods of "tan", and "can". Thus, the structure of the lexical neighborhoods of pediatric CI users may be quite different from the structure of the lexical neighborhoods of a listener with normal hearing.

Although the lexical neighborhoods of a pediatric CI user may be broader than the lexical neighborhoods of a listener with normal hearing, the average size of the receptive vocabulary of the pediatric CI user or deaf child is considerably smaller than the average size of the receptive vocabulary of a group of children of the same age with normal hearing (Geers & Moog, 1994; Miyamoto et al., 1992). A smaller vocabulary should lead, therefore, to lexical neighborhoods that are, in general, less dense (contain fewer words) than lexical neighborhoods for listeners with normal hearing and normal vocabulary development.

At the present, we are conducting an analysis of the errors made by pediatric CI users with two and three years of implant use on open-set tests of word recognition (Meyer, Wright, Chin, & Pisoni, 1998). This analysis should give us a better understanding of the neighborhood structure of the children with CIs and provide insight into how these similarity neighborhoods change with increased implant use. As the ability of a pediatric CI user to make fine distinctions between the different acoustic-phonetic properties of words improves, the listener's lexical neighborhood structure for a particular word should become more refined and decrease in size. Preliminary analyses of the error data provide support for this prediction.

Summary and Conclusions

The PBK Test has been used for almost 50 years to assess the open-set speech perception skills of young children. Originally, Haskins generated four lists of words, but after testing the intelligibility of her lists at different S/N ratios with a group of normal-hearing adult listeners, she concluded that one of the word lists (PBK List 2) was easier (i.e., "more audible") than the other three lists and should not be used clinically. In the present study, we examined the lexical properties of all of the words in the four lists of the PBK Test to determine if any differences could be found in the similarity spaces of the words used on these lists in terms of measures of word frequency and lexical density. The results of our computational analysis demonstrate that the words on List 2 are of higher frequency and come from less dense neighborhoods than the words in the remaining three PBK lists.

The results of this analysis lend further support to the hypotheses set forth by Luce and his colleagues in describing NAM (Luce, 1986; Luce & Pisoni, in press; Luce, Pisoni & Goldinger, 1990) that words are recognized relationally in the context of other words. Word frequency and lexical density control the recognition process for words presented in isolation.

References

American National Standards Institute. (1989). Specifications for audiometers. (ANSI S3.6-1989). New York: ANSI.

Carney, A. E., Osberger, M. J., Miyamoto, R. T., Karasek, A., Dettman, D. L., & Johnson, D. L. (1991). Speech perception along a continuum: From hearing aids to cochlear implants. In Feigin, J. A. & Stelmachowicz, P. G. (Eds.) *Pediatric Amplification*. (pp. 93-113). Omaha, NE: Boystown National Research Hospital.

Dunn, L., & Dunn, L. (1981). Peabody picture vocabulary test - revised. Circle Pines, MN: American Guidance.

Egan, J. P. (1948). Articulation testing methods. *Laryngoscope, 58*, 955-991.

Fryauf-Bertschy, H., Tyler, R. S., Kelsay, D. M., & Gantz, B. J. (1992). Performance over time of congenitally deaf and postlingually deafened children using a multichannel cochlear implant. *Journal of Speech and Hearing Research, 35*, 913-920.

Fryauf-Bertschy, H., Tyler, R. S., Kelsay, D. M.,Gantz, B. J., & Woodworth, G. G. (1997). Cochlear implant use by prelingually deafened children: The influences of age at implant and length of device use. *Journal of Speech, Language, and Hearing Research, 40*, 183-199.

Gantz, B. J., Tyler, R. S., Tye-Murray, N., Fryauf-Bertschy, H. (1994). Long term results of multichannel cochlear implants in congenitally deaf children. In Hochmair-Desoyer, I. J. & Hochmair, E. S. (Eds.) *Advances in Cochlear Implants* (pp. 528-533). Vienna, Austria: Manz.

Gantz, B. J., Tyler, R. S., Woodworth, G. G., Tye-Murray, N., & Fryauf-Bertschy, H. (1994). Results of multichannel cochlear implants in congenital and acquired prelingual deafness in children: Five-year follow-up. *American Journal of Otology, 12*, 1-7.

Geers, A. E., & Moog, J. S. (1994). Spoken language results: Vocabulary, syntax, and communication. *The Volta Review, 96*, 131-148.

Greenberg, J. H., & Jenkins, J. J. (1963). Studies in the psychological correlates of the sound system of American English. *Word, 20*, 157-177.

Kirk, K. I., Pisoni, D. B., & Osberger, M. J. (1995). Lexical effects on spoken word recognition by pediatric cochlear implant users. *Ear & Hearing, 16*, 470-481.

Kucera, F., & Francis, W. (1967). *Computational Analysis of Present-Day American English.* Providence, RI: Brown University Press.

Lane, H. (1995). Acquisition of speech perception ability in prelingually DEAF children with a multichannel cochlear implant. Letter to the Editor. *American Journal of Otology, 16*, 393-399.

Luce, P. A., & Pisoni, D. B. (in press). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing.*

Luce, P. A., Pisoni, D. B., & Goldinger, S. D. (1990) Similarity neighborhoods of spoken words. In G. T. M. Altmann, (Ed.), *Cognitive Models of Speech Processing: Psycholinguistic and Computational Perspectives* (pp. 122-147). Cambridge, MA: MIT Press.

MacWhinney, B., & Snow, C. (1985). The child language data exchange system. *Journal of Child Language, 12*, 271-296.

Miyamoto, R. T., Osberger, M. J., Robbins, A. M., Myres, W. A., Kessler, K., & Pope, M. (1992). Longitudinal evaluation of communication skills of children with single- or multichannel cochlear implants. *American Journal of Otology, 13*, 215-222.

Miyamoto, R. T., Osberger, M. J., Robbins, A. M., Myres, W. A., & Kessler, K. (1993). Prelingually deafened children's performance with the Nucleus multichannel cochlear implant. *American Journal of Otology, 14*, 437-445.

Miyamoto, R. T., Osberger, M. J., Todd, S. L., Robbins, A. M., Stroer, M. A., Zimmerman-Phillips, S., & Carney, A. E. (1994). Variables affecting implant performance in children. *Laryngoscope, 104*, 1120-1124.

Osberger, M. J., Miyamoto, R. T., Zimmerman-Phillips, S., Kemink, J. L., Stroer, B. S., Firszt, J. B., & Novak, M. A. (1991). Independent evaluation of the speech perception abilities of children with the Nucleus 22-channel cochlear implant system. *Ear & Hearing, 12*, S66-S80.

Pisoni, D. B., Nusbaum, H. C., Luce, P. A., & Slowiaczek, L. M. (1985). Speech perception, word recognition and the structure of the lexicon. *Speech Communication, 4*, 75-95.

362

Pollack, I., Rubenstein, H., & Decker, L. (1959). Intelligibility of known and unknown message sets. *Journal of the Acoustical Society of America, 31*, 273-279.

Staller, S. J., Beiter, A. L., Brimacombe, J. A., Mecklenburg, D. J., & Arndt, P. (1991). Pediatric performance with the Nucleus 22-channel cochlear implant system. *American Journal of Otology, 12*, S126-S136.

Staller, S. J., Dowell, R. C., Beiter, A. L., Brimacombe, J. A. (1991). Perceptual abilities of children with the Nucleus 22-channel cochlear implant. *Ear & Hearing, 12*, S34-S47.

Svirsky, M. A. (1996). Speech production and language development in pediatric cochlear implant users. *Journal of the Acoustical Society of America, 99*, 2570.

Waltzman, S., Cohen, N., Gomolin, R., Ozdamar, S., Shapiro, W., & Hoffman, R. (1995). Effects of short-term deafness in young children implanted with the Nucleus cochlear prosthesis. *Annals of Otology, Rhinology, & Laryngology, 104*, 341-342.

Waltzman, S. B., Cohen, N. L., Gomolin, R. H., Shapiro, W. A., Ozdamar, S. R., & Hoffman, R. A. (1994). Long-term results of early cochlear implantation in congenitally and prelingually deafened children. *American Journal of Otology, 12*, 9-13.

Young, L. L., Dudley, B., & Gunter, M. B. (1982). Thresholds and psychometric functions of the individual spondaic words. *Journal of Speech and Hearing Research, 25*, 586-593.

II. Short Reports & Work-in-Progress

Cognitive Factors and Cochlear Implants: An Overview of the Role of Perception, Attention, Learning and Memory in Speech Perception[1]

David B. Pisoni[2]

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405

---

[1] This research is supported by NIH NIDCD Research Grants DC00064, DC00423 and DC00111 to Indiana University. Text of an invited paper presented at the Vth International Cochlear Implant Conference, May 1-3, 1997, New York, NY.

[2] Also DeVault Otologic Research Laboratory, Department of Otolaryngology–Head and Neck Surgery, Indiana University School of Medicine, Indianapolis, IN.

Cognitive Factors and Cochlear Implants: An Overview of the Role of Perception, Attention, Learning and Memory in Speech Perception

Abstract. Over the last few years, there has been increased interest in studying the cognitive factors that affect speech perception performance of cochlear implant patients. In this paper, I provide a brief theoretical overview of the information processing approach to cognition and discuss the role of perceptual learning, attention and memory in speech perception and spoken language processing. Directions for future research on information processing issues are discussed with the goal of predicting success with a cochlear implant from a set of cognitive measures of performance.

Introduction

We are now beginning to see several important changes in the direction and nature of research on cochlear implants, particularly research on very young prelinguistically deaf children who have received cochlear implants. As cochlear implants and their speech processing strategies improve, more and more deaf children are able to derive greater benefit from their implants. Many of these children display substantial gains in speech perception, word recognition and language development (Fryauf-Bertschy, Tyler, Kelsay, & Gantz, 1992; Fryauf-Bertschy, Tyler, Kelsay, Gantz, & Woodworth, 1997; Miyamoto, Kirk, Robbins, Todd, Riley, & Pisoni, 1997; Miyamoto, Svirsky, & Robbins, 1997; Waltzman et al., 1997). And, many prelinguistically deafened children with CI's somehow learn how to produce intelligible speech within one year of implantation and appear to be well on their way to acquiring a grammar of spoken language via their CI (Miyamoto, Svirsky et al., 1997; Robbins, Svirsky, & Kirk, in press). How do they accomplish this difficult task?

As performance levels on standardized audiological tests continue to improve, a number of researchers have turned their efforts to gaining a better understanding of several more basic questions that surround how cochlear implants function to facilitate cognitive and linguistic development in deaf children (Kirk, Pisoni, & Miyamoto, in press; Kirk, Pisoni, & Osberger, 1995; Miyamoto, Svirsky et al., 1997; Robbins, Svirsky, & Kirk, in press). These new questions deal with a variety of issues involving the perception of speech and spoken language understanding. More generally, the interest is starting to shift to questions of how deaf children encode and process information using a cochlear implant. Many of these questions concern fundamental issues of human information processing and involve topics such as perceptual learning, memory, attention and language processing, research areas that have traditionally been in the mainstream of Cognitive Psychology (Ashcraft, 1989; Atkinson & Shiffrin, 1968; Crowder, 1976; Haber, 1969) and Cognitive Science (Gardner, 1985).

Much of the past research on CI's has been concerned with questions of assessment and device efficacy using outcome measures that were based on traditional audiological criteria. These measures included a variety of hearing tests, speech discrimination, word recognition and comprehension tests, as well as some standardized vocabulary and language assessments. The major focus of this research over the last 10-15 years has been concerned with the study of demographic variables as predictors of these outcome measures. The available evidence demonstrates that age at onset of deafness, length of deprivation and age at implantation play substantial roles in predicting many of the standard outcome measures

(Fryauf-Bertschy et al., 1997; Kirk et al., in press; Osberger, Todd, Berry, Robbins, & Miyamoto, 1991; Staller, Pelter, Brimacombe, Mecklenberg, & Arndt, 1991; Waltzman et al., 1994, 1997). What happens if you eliminate the demographic variables? What is left over to study? We suggest there are a number of "process" variables or factors that are related to learning, memory, attention and language processing that have been ignored over the years. We feel that these particular areas of research are critical to gaining new insights into how children acquire language through a cochlear implant and explaining the enormous individual differences among prelingually deaf children with cochlear implants.

The child's early sensory experience has also been shown to have a significant role in predicting outcome measures. It should not come as a surprise to anyone that deaf children from "oral-only" programs do consistently better on auditory-based tests of speech perception and language performance than deaf children from total communication (or TC, i.e., manually-coded language plus speech) programs (Kirk, 1996; Miyamoto, Kirk et al., 1997; Robbins, Kirk, Osberger, & Ertmer, 1995). The study of demographics and the focus on traditional audiological outcome measures in these children are only a small part of the story of what is actually going on. To gain a better understanding of what these children are learning via their implant, it is necessary to approach this problem from an entirely different theoretical perspective and to look more closely at the content and flow of information and how it changes over time and study the underlying processes.

Little, if any, of the previous research on cochlear implants has been concerned with studying what the children are learning via their implant, how they are going about the process of acquiring a grammar from the ambient language or how they are able to develop both receptive and expressive language abilities. Moreover, until recently there have been very few attempts to study the language development of children with CI's and compare their linguistic knowledge and performance with normal hearing children or with other hearing-impaired children (Miyamoto, Svirsky et al., 1997; Robbins & Kirk, 1996). These are important questions that go well beyond surface issues of assessment, device efficacy, or simply predicting outcome measures; they are fundamental questions that deal with the "effectiveness" of cochlear implants outside the special conditions of the clinic or the research laboratory. The major emphasis on assessment-based clinical research and the prediction of outcome measures is changing now, and there are several papers at this meeting that report new findings on some of these important new questions (Kirk et al., in press; Pisoni, Svirsky, Kirk, & Miyamoto, this volume; Robbins et al., in press; Zwolan et al., 1997).

In order to explore some of these new research questions and to move beyond the study of demographics and the issues surrounding assessment and prediction of outcome measures, it is necessary to look to other allied disciplines such as Cognitive Psychology (Haber, 1969; Neiser, 1967; Reitman, 1965) and Cognitive Science (Gardner, 1985). New experimental methods and techniques must be used to study the emergence of these fundamental underlying cognitive and neural processes and how these processes change over time after implantation. Fortunately, many useful experimental procedures have already been developed by cognitive psychologists to study perception, attention, learning and memory within the framework of human information processing (Haber, 1969; Lachman, Lachman, & Butterfield, 1979; Neisser, 1967). This approach has also provided a variety of conceptual tools for thinking about the fundamental structures and processes involved in cognitive activity and the underlying psychological phenomena (Lindsay & Norman, 1977; Reitman, 1965).

"Information processing" is a label for a general approach to the study of complex psychological processes such as perception, cognition and thought (Haber, 1969; Neisser, 1967). Information processing theories are concerned with an analysis of "central processes" of large complex systems (such as human cognition) used in visual object recognition, perceptual learning and memory, speech perception, and various aspects of language processing such as comprehension or speech production. A common goal of this approach is to examine the representations, elementary psychological processes and cognitive structures used in these cognitive activities and to trace out the time course of these processing operations (Haber, 1969; Lachman et al., 1979; Sternberg, 1966, 1969).

In the sections below, I first give a very brief overview of the major theoretical assumptions of this approach to cognition and to areas of research such as perception, attention, learning and memory. Then, I will examine several new directions for future research on cochlear implants that are motivated by the major assumptions of the information processing framework. I believe it may now be possible to understand and explain the large individual differences observed in children and adults with cochlear implants by studying the psychological and cognitive factors and the component subsystems used in perception, attention, learning and memory. This is one of many problems that can now be approached with some confidence within this theoretical framework.

Overview of the Information Processing Approach

Assumption I: Perception is Not Immediate

One of the fundamental principles of information processing theory is that sensation, perception, memory, thought and other complex activities like language and problem solving should be viewed as representing a continuum of cognitive processing (Haber, 1969; Neisser, 1967). These activities are assumed to be mutually interdependent and cannot be divided up into separate subsystems. Furthermore, an analysis of one subsystem, such as perception, cannot take place without an appreciation and awareness of the contribution of the other major subsystems, such as memory, attention or learning (Haber, 1969; Neisser, 1967; Reitman, 1965).

The information processing approach to cognition also assumes that processing activity goes through several successive stages of analysis. One goal of information processing theory is to specify the component operations that occur between the presentation of a stimulus and the response of the observer. The processing stages between input and output are typically represented by a flow chart with structures and processes organized in a block diagram. The flow of information is marked by arrows connecting these structures (Haber, 1969).

These hypothesized processing stages also take time. It is assumed that processing times reflect distinct operations that occur at each stage (Baddeley, 1986). By looking at the correlations between the contents of the stimulus and the contents of the observer's response at various times after stimulation, some insights can be gained about the flow of information within the system and the nature of the operations being carried out at each stage of processing (Haber, 1969; Neisser, 1967; Reitman, 1965).

Finally, the information processing approach assumes that psychological processes such as sensation, perception, attention, learning and memory are organized hierarchically. More complex cognitive processes which occur later in the flow of information are critically dependent on earlier more elementary psychological processes (Atkinson & Shiffrin, 1968; Schneider & Shiffrin, 1977; Shiffrin & Schneider, 1977).

Assumption II: Capacity Limitations on Processing Information

A second principle of information processing theory is that the human observer has finite information processing capabilities and displays severe channel capacity limitations in a variety of tasks (Broadbent, 1958; Miller, 1956; Waugh & Norman, 1965). Years ago it was believed that the nervous system was not large enough to maintain all aspects of stimulation permanently and that raw sensory information needed to be transformed, reduced and recoded into a more efficient symbolic form for storage in memory (Neisser, 1967). A central problem in information processing theory is identifying the "locus" of where recoding takes place in the processing system and describing the nature of these processing operations (Kahnman, 1973; Shiffrin, 1988). Thus, research in cognitive psychology on topics such as selective attention (Cherry, 1953; Lindsay & Norman, 1977; Shiffrin, 1988) and immediate memory span (Baddeley & Hitch, 1974; Miller, 1956; Waugh & Norman, 1965) demonstrate that not all aspects of the stimulus environment are encoded or processed by the nervous system or stored in permanent memory for later retrieval (Cowan, 1988; Navon & Gopher, 1979). Understanding the process of "information reduction" and "recoding" by the nervous system has been a long-standing problem that cuts across several domains including perception, attention, learning and categorization (Baddeley, 1990; Posner, 1969).

Assumption III: Commonality of Perception and Memory

A third principle of information processing theory is that all aspects of cognitive activity—ranging from sensation and perception to learning and thought—involve some kind of storage or memory system that preserves selected aspects of the initial sensory stimulation (Atkinson & Shiffrin, 1968; Lindsay & Norman, 1977; Posner, 1969; Waugh & Norman, 1965). Thus, the nature of the neural representations of the stimulus in memory and the organization of this knowledge is a fundamental problem in all information processing analyses. The operations involved in encoding, rehearsal, storage and retrieval of information occur at all stages of processing (Atkinson & Shiffrin, 1968). This theoretical approach also assumes that it is not possible to separate the processes that support perception from those that support memory. The two processes are mutually dependent. Memory is therefore one of the central problems to be studied in understanding psychological activities within information processing theory. Encoding, storage, retrieval and rehearsal of the stimulus input all take place within this common subsystem (Atkinson & Shiffrin, 1968; Baddeley, 1986, 1990; Baddeley & Hitch, 1974; Craik & Lockhart, 1972; Posner & Mitchell, 1967).

Goals of Information Processing Approach

The information processing approach is concerned primarily with the "central" cognitive processes used in large systems to carry out complex activities such as perception, learning, memory, language processing and problem solving. The goals of this approach follow from the three principles described earlier. First, information processing theory focuses research on describing the sequence of operations, or "stages of processing," used in a particular task. Second, this approach is concerned with identifying the "locus" of capacity limitations in processing information. Third, this approach attempts to trace the "time-course" of perceptual processing from the stimulus input that impinges on the nervous system to the observer's overt response. Fourth, researchers working within the information processing approach attempt to construct process models and computer simulations of various subsystems in order to formalize and make precise quantitative statements about their performance (Hunt, 1978; Lindsay & Norman, 1977; McClelland & Elman, 1986; McClelland & Rumelhart, 1986; Rumelhart & McClelland, 1986). These models provide detailed explanations of phenomena and make explicit predictions. Finally, over the last few years, researchers working within this general framework have tried to establish the neural plausibility of their models (e.g., McClellan & Rumelhart, 1986; Rumelhart & McClelland, 1986). In the past, most cognitive psychologists expressed little, if any, interest in brain modeling. Their concern was with the flow of abstract symbolic information within the information processing system and with the correlations between stimulus and response at various points in time after stimulation terminated. This situation is changing now as new concepts and techniques from neural networks and connectionist models become available to permit the construction of neurally-inspired models that have some relationship to what is currently known about the nervous system and brain function.

Methodology of Converging Operations

An important methodological principle of the information processing approach is the emphasis on "converging operations" (Garner, Hake, & Eriksen, 1956; Haber, 1969). The idea here is to obtain data on the flow and content of information within the processing system using a wide variety of experimental techniques and research designs and to search for commonalties across different tasks. It is also assumed that the processing activities at different stages can be revealed by two general measures of an observer's behavior: the accuracy of performance and the time required to perform a given task (Hunt, 1978; Lachman et al., 1979; Sternberg, 1966, 1969). Both measures have a long history in the field of experimental psychology and both measures have provided a great deal of valuable information about the component elementary psychological operations used to carry out a particular task. Researchers have also been interested in examining the incorrect responses of subjects in a variety of experimental paradigms in order to analyze error patterns. Errors are often quite systematic and provide new insights into the way human perceivers use partial stimulus information to structure their responses. All of these measures provide information about the underlying cognitive processes being studied.

Some New Research Directions on Cochlear Implants in Children

Language Development vs. Hearing

The bulk of research on cochlear implants has been carried out by audiologists and hearing scientists who have been concerned with the sensory coding of speech by the peripheral nervous system. Only recently have researchers begun to examine the effects of cochlear implants on specific aspects of language development. One very important area of research on language development concerns the child's phonological system which encodes and represents the inventory of sounds the child has acquired and the sound contrasts of the ambient language (Chin, Pisoni, & Svec, 1993).

In order for a child to produce intelligible speech, he must have an organized system for encoding and representing sound patterns in memory, a phonology, and a set of procedures to translate these phonological representations and rules into sensory-motor commands and gestures that can control the vocal tract and articulators in speech production. It is generally assumed that the child has one common phonological system of representation that is used for both speech perception and speech production. Unfortunately, very little is currently known about the phonological systems of prelinguistic deaf children with cochlear implants. Research on this topic has not generated much interest among audiologists, who are concerned primarily with hearing and the peripheral auditory system. Only recently have clinical phonologists begun to seriously study this problem in deaf children with cochlear implants (Chin & Kirk, in press; Chin et al., 1993). These findings indicate that deaf children with cochlear implants display several commonalties with normal-hearing children in terms of their inventory of sounds and the patterning of these sounds in production (Chin et al., 1993).

The same situation is also true for the study of spoken word recognition and lexical access, two subcomponents of the language comprehension system. Little, if any, research has been done on the organization of the child's developing lexicon or the nature of the lexical representations of words that are constructed by deaf children with cochlear implants (Kirk, 1996; Kirk et al., 1995). Some recent evidence suggests that deaf children with cochlear implants perceive and represent spoken words in terms of broad phonetic categories or functional equivalence classes that reflect their inability to reliably discriminate fine phonetic differences in place and voicing (Pisoni, Svirsky, Kirk, & Miyamoto, this volume). Difficulty in perceiving and encoding phonetic distinctions among sound patterns would in all likelihood influence the organization and structural arrangement of words in the lexicon (see Logan, 1992), and would no doubt produce parallel changes in speech and language production as well. These are two topics that are being explored in our research laboratory.

At the present time, very little research has been carried out on morphological and syntactic development in deaf children with cochlear implants. This is a critical area of language development that needs to be investigated in greater depth in order to determine whether children can acquire abstract linguistic knowledge about the grammar of the target language through a processing device that presents their nervous system with a highly degraded and impoverished electrical signal. The key question here is whether deaf children can acquire the full range of morphological contrasts and structural regularities of English when they can represent the sound contrasts of the language only in terms of broad manner classes. What kind of morphological system will these children actually come up with and how is it different from the system normal-hearing children develop? This is obviously an interesting and important research problem because it deals with the interface between phonology and syntax.

Perception and Production

Historically, the fields of speech perception and production have developed independently of each other. In deaf children who have received cochlear implants, it may be necessary to study the development of both processes together in order to gain insights into the underlying linguistic system of the child. Recent findings have demonstrated unusually high correlations between open-set word recognition scores and measures of speech intelligibility (Pisoni et al., this volume). We need to learn more about the child's acquisition and use of phonological information. Specifically, we want to know whether their systems reflect language universals or a coding limitation of the cochlear implant regarding certain phonetic features of the speech signal like place and voicing. Recent findings suggest a common source of variance underlying word recognition and speech intelligibility that involves the phonological representations of words and the mapping of sound patterns onto meanings in memory (Pisoni et al., this volume). Whatever linguistic skills or abilities these children employ in recognizing words in isolation (i.e., without any context or retrieval cues) also appear to be recruited in speech production. The pattern of intercorrelations we have found in our recent analysis of the "Stars" suggests a common underlying representational system for phonological knowledge in memory.

Multimodal Speech Perception

Although many researchers and theorists have traditionally viewed speech perception and spoken language processing as purely acoustic/auditory operations, recent findings on multimodal perception (Massaro, 1998) have provided many reliable demonstrations of the visual/optical correlates for speech perception as represented in the dynamic changes in the talker's face and lips. This topic has many implications for the hearing-impaired because these perceivers often rely heavily on information in the optical display of a talker's face as an aid to speech perception. If speech is viewed within the theoretical framework of event perception as a perceptual system having both acoustic and optical correlates (Auer & Bernstein, 1997; Gaver, 1993), then our view of the task confronting the perceiver must be modified accordingly to fully acknowledge the multimodal properties of speech and the lawful relations between auditory and visual speech cues. Normal-hearing listeners often have difficulty dissociating these two sensory inputs and respond in ways suggesting an integrated perceptual pattern.

These observations about multimodal speech perception are relevant to several recent findings showing interference and inhibition effects of manual communication skills in TC children who are learning oral language via their cochlear implant. In these TC children, knowledge and use of sign language apparently competes with the dominant mode of processing speech via the auditory/phonetic modality. The differences in modality between sign language and speech apparently prevents these TC children from integrating common information across sensory modalities, therefore increasing the processing load on working memory which is assumed to play a major role in language comprehension and word recognition (Baddeley, Gathercole & Papagno, 1998).

Perceptual and Cognitive Development

Many basic questions about perceptual and cognitive development have not yet been studied in deaf children with cochlear implants. At this time, we know very little about the perceptual learning abilities of these children or how auditory and visual attention is shaped and modified by awareness of sound and perception of speech after long periods of sensory deprivation (Lenneberg, 1967). Almost no research has been done on categorization or concept learning in these children. Similarly, we know almost nothing about their working memory systems, a key factor in acquiring new words and producing spoken language using phonological knowledge previously stored in memory (Baddeley et al., 1998). Finally, we currently have no systematic knowledge or information about the metalinguistic abilities of deaf children with cochlear implants. It seems reasonable at this point to wonder if the children who have acquired some rudimentary language skills via their implant also have explicit metalinguistic abilities to reason about and communicate about spoken language as an abstract system. Results from the reading literature suggest that metalinguistic awareness is a strong predictor of early reading success. Is the same relationship true of deaf children with cochlear implants?

Looking at the "Stars"—Studies of the Exceptional Users of Cochlear Implants

The published literature on cochlear implants has consistently reported large individual differences among users. Some prelinguistically deaf children do exceptionally well with their implants, and go on to acquire spoken language and produce intelligible speech. Other children, however, develop only an awareness of sound and never appear to acquire language or produce intelligible speech to the same degree or proficiency as the exceptionally good users. We are now just beginning to examine the performance of the exceptionally good users of cochlear implants, the so-called "Stars," on a variety of behavioral measures including open- and closed-set speech perception tests, word recognition, and vocabulary tests, as well as expressive and receptive language development (Pisoni et al., this volume). Our analyses of the intercorrelations of these measures suggest that the "Stars" have developed a representational system, that is, a phonology and a lexicon for mapping sounds onto meanings. Their exceptionally good abilities in recognizing words spoken in isolation are not restricted to only open-set word recognition tasks. The "Stars" display very good performance on several other tasks, all of which apparently require access to and use of words stored in the lexicon. The pattern of intercorrelations with other behavioral measures was extremely strong for the Reynell expressive and receptive language scales (Pisoni et al., this volume). These recent findings show that the children who do well on open-set speech perception and word recognition skills also do well on other language measures. Finally, one of the most important discoveries from our analysis were the high intercorrelations of the open-set word recognition tests with measures of speech intelligibility. Whatever skills, abilities or processes children use to recognize isolated words in an open-set format, these same processes also appear to be recruited in speech production when the child has to access sensory-motor patterns in order to repeat back spoken words.

Assessment-Based vs. Theory-Driven Research

The primary focus of most of the past research on cochlear implants in children has been on device efficacy and predicting outcome measures using standardized audiological tests. Because of these goals, researchers have tended to concentrate on the study of demographic variables as predictors of success with a cochlear implant. Until recently, these were the only independent variables that were included in the research designs used to study the performance of children with cochlear implants. This is not too surprising given the theoretical orientation and research background of most of the investigators who work on cochlear implants. Audiologists are trained in hearing assessment and traditionally they have had very little interest or motivation in underlying theory. In fact, one could argue that the field of audiology is, for the most part, atheoretical in its approach to hearing and speech perception. The situation is now changing in several respects as the research questions focus in on a variety of new issues surrounding what the child is learning via the cochlear implant and how the cochlear implant works in a functional way.

As we noted earlier, these are research questions that deal with psychological processing activities underlying the actual use of the cochlear implant. Fundamental questions about perception, learning, memory, attention and language which lie outside the domain of clinical audiology or hearing science, can all be approached within the framework of information processing theory because of its concern with describing the underlying psychological processes and mechanisms that intervene between stimulus input and response output. Viewed within this broader theoretical context, many of the difficult "central processing" issues surrounding topics such as individual differences, the time-course of language development, and the relations between speech perception and speech production can now be approached using a variety of new concepts and experimental techniques. The emphasis on demographics no longer has to be the primary focus of research on deaf children with cochlear implants. There are many more important new questions to study.

The shift from "assessment-based" research to "hypothesis testing" and theory-driven research represents a natural progression as researchers move from simple description and device "efficacy" questions to explanation, prediction, and "effectiveness" issues. Fundamentally, we want to know what deaf children are learning via their cochlear implant and how they manage to accomplish this task. Answers to these basic questions about underlying psychological process may have broad implications for new approaches to processor design, aural rehabilitation, and decision making with prelinguistically deaf children. The findings that some deaf children with cochlear implants can perceive speech and produce spoken language is very encouraging because it demonstrates device "efficacy." That is, cochlear implants work with some deaf children, and these children appear to acquire spoken language in spite of using a highly degraded and impoverished electrical signal. However, we do not know how this is accomplished in the exceptionally good children like the "Stars" nor do we know why other children have more difficulty in reaching these important goals. If we had some better ideas and specific hypotheses about what psychological processes and mechanisms were responsible for the exceptionally good performance of the "Stars," we might be able develop new intervention techniques to accelerate and improve the perceptual learning and language development of the "average" user of a cochlear implant. It is very unlikely that changes like this would ever come about by continuing to do descriptive assessment-based research with these children using the traditional measurement techniques from hearing science and clinical audiology. What is needed now is an integrated theoretical framework for studying perceptual learning in these children and relating these findings to performance on speech and language tests.

We believe the information processing approach to complex psychological activities has a great deal to offer at this time. We are encouraged already by several new findings on the "Stars" who display exceptionally good performance on a wide variety of behavioral tests of speech perception and language processing. The "Stars" no longer need to be viewed as anomalies, but may instead provide deep insights into the underlying cognitive processes that are responsible for their superior performance across many different tests. This theoretical framework should also provide us with new ways to study and understand the time-course of perceptual and cognitive development and the interrelations between speech perception and production in these children. Hopefully, these new research directions will help us to understand the role of the environment and the effects of early experience on language development during the critical period when the child's nervous system is still amenable to change.

References

Ashcraft, M. H. (Ed.). (1989). *Human memory and cognition*. Glenview, IL: Scott, Foresman and Company.

Auer, E.T., & Bernstein, L.E. (1997). Speechreading and the structure of the lexicon: Computationally modeling the effects of reduced phonetic distinctiveness and lexical uniqueness. *Journal of the Acoustical Society of America*, **102**, 3704-3709.

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In K. W. Spence & J. T Spence (Eds.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 2, pp. 89-195). New York: Academic Press.

Baddeley, A. D. (1986). *Working memory*. New York: Oxford University Press.

Baddeley, A. D. (1990). *Human memory: Theory and practice*. Boston: Allyn & Bacon.

Baddeley, A., Gathercole, S. & Papagno, C. (1998). The phonological loop as a language learning device. *Psychological Review*, **105**, 158-173.

Baddeley, A. D., & Hitch, G. J. (1974). Working memory. In G. A. Bower (Ed.), *The psychology of learning and motivation* (vol. 8, pp. 47-90). New York: Academic Press.

Broadbent, D. E. (1958). *Perception and communication*. Oxford: Pergamon Press.

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and two ears. *Journal of the Acoustical Society of America*, **25**, 975-979.

Chin, S.B., & Kirk, K. I. (in press). Consonant feature production by children with multichannel cochlear implants, hearing aids, and tactile aids. In S. Waltzman and N. Cohen (Eds.), *Proceedings of the Vth International Cochlear Implant Conference*. New York: Thieme Medical Publishers.

374

Chin, S.B., Pisoni, D.B., & Svec, W.R. (1993). An emerging phonetic-phonological system two years post-cochlear implant: A preliminary linguistic description. In *Research on Spoken Language Processing Progress Report No. 19* (pp. 253-270). Bloomington, IN: Speech Research Laboratory, Indiana University.

Cowan, N. (1988). Evolving conceptions of memory storage, selective attention, and their mutual constraints within the human information processing system. *Psychological Bulletin*, **104**, 163-191.

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, **11**, 671-684.

Crowder, R. G. (1976). *Principles of learning and memory*. Hillsdale, NJ: Erlbaum.

Fryauf-Bertschy, H., Tyler, R.S., Kelsay, D. & Gantz, B.J. (1992). Performance over time of congenitally and postlingually deafened children using a multichannel cochlear implant. *J Speech Hear Res*, **35**, 913-920.

Fryauf-Bertschy, H., Tyler, R.S., Kelsay, D., Gantz, B.J. & Woodworth, G.G. (1997). Cochlear implant use by prelingually deafened children: The influences of age at implant and length of device use. *Journal of Speech, Language, and Hearing Research*, **40**, 183-199.

Gardner, H. (1985). *The mind's new science: A history of the cognitive revolution*. New York: Basic Books.

Garner, W. R., Hake, H. W., & Eriksen, C. W. (1956). Operationism and the concept of perception. *Psychological Review*, **63**, 317-329.

Gaver, W.W. (1993). What in the world do we hear? An ecological approach to auditory event perception. *Ecological Psychology*, **5**, 1-29.

Haber, R. N. (Ed.). (1969). *Information-processing approaches to visual perception*. New York: Holt, Rinehart and Winston.

Hunt, E. B. (1978). Mechanics of verbal ability. *Psychological Review*, **85**, 109-130.

Kahnman, D. (1973). *Attention and effort*. Englewood Cliffs, NJ: Prentice Hall.

Kirk , K. I. (1996). Lexical discrimination and age at implantation. A first report. Paper presented at the 131st meeting of the Acoustical Society of America, Indianapolis, IN.

Kirk, K.I., Pisoni, D.B. & Miyamoto, R.T. (in press). Lexical discrimination by children with cochlear implants: Effects of age at implantation and communication mode. In S. Waltzman & N. Cohen (Eds.) *Proceedings of the Vth International Cochlear Implant Conference*. New York: Thieme Medical Publishers.

Kirk, K.I., Pisoni, D.B. & Osberger, M.J. (1995). Lexical effect on spoken word recognition by pediatric cochlear implant users. *Ear & Hearing*, **16**, 470-481.

Lachman, R., Lachman, J. L., & Butterfield, E. C. (1979). *Cognitive psychology and information processing: An introduction*. Hillsdale, NJ: Erlbaum.

Lenneberg, E. (1967). *Biological foundations of language*. New York: Wiley.

Lindsay, P. H., & Norman, D. A. (1977). *Human information processing: An introduction to psychology*. New York: Academic Press.

Logan, J. S. (1992). *A computational analysis of young children's lexicons* (Research on Spoken Language Processing, Technical Report No. 8). Bloomington, IN: Speech Research Laboratory, Indiana University.

Massaro, D.W. (1998). *Perceiving talking faces: From speech perception to a behavioral principle*. Cambridge, MA: The MIT Press.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, **63**, 81-97.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, **18**, 1-86.

McClelland, J. L., & Rumelhart, D. E. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition. Vol. 2: Psychological and biological models*. Cambridge, MA: Bradford.

Miyamoto, R. T., Kirk, K. I., Robbins, A. M., Todd, S., Riley, A., & Pisoni, D. B. (1997). Speech perception and speech intelligibility in children with multichannel cochlear implants. In I. Honjo & H. Takahashi (Eds.), *Cochlear implant and related sciences update*. Advances in Otorhinolaryngology (pp. 198-203). Basel, Switzerland: Karger.

Miyamoto, R. T., Svirsky, M. A., & Robbins, A. M. (1997). Enhancement of expressive language in prelingually deaf children with cochlear implants. *Acta Otolaryngologica*, **117**, 154-157.

Navon, D., & Gopher, D. (1979). On the economy of the human-processing system. *Psychological Review*, **86**, 214-255.

Neisser, U. (1967). *Cognitive psychology*. New York: Appleton-Century-Crofts.

Osberger, M.J., Todd, S.L., Berry, S.W., Robbins, A.M. & Miyamoto, R.T. (1991). Effect of age of onset of deafness on children's speech perception abilities with a cochlear implant. *Ann Otol Rhinol Laryngol*, **100**, 883-888.

Pisoni, D.B., Svirsky, M.A., Kirk, K.I., & Miyamoto, R.T. (this volume). Looking at the "starts": A first report on the intercorrelations among measures of speech perception, intelligibility and language in pediatric cochlear implant users. In *Research on Spoken Language Processing Progress Report No. 21* (pp. 51-92). Bloomington, IN: Speech Research Laboratory, Indiana University.

Posner, M. I. (1969). Abstraction and the process of recognition. In G. H. Bower & J. T. Spence (Eds.), *The psychology of learning and motivation* (Vol. 3, pp. 43-100). New York: Academic Press, pp. 43-100.

Posner, M. I., & Mitchell, R. F. (1967). Chronometric analysis of classification. *Psychological Review*, **74**, 392-409.

Reitman, W. R. (1965). *Cognition and thought: An information processing approach*. New York: John Wiley & Sons.

376

Robbins, A. & Kirk, K. I., (1996). Speech perception assessment and performance in pediatric cochlear implant users. *Seminars in Hearing*, 17, 353-369.

Robbins, A. M., Kirk, K. I., Osberger, M. J., & Ertmer, D. J. (1995). Speech intelligibility of implanted children. *Annals of Otology, Rhinology, & Laryngology*, **104**, 399-401.

Robbins, A., Svirsky, M., Kirk, K. I. (In Press). Children with implants can speak, but can they communicate? *Otolaryngology-Head & Neck Surgery*.

Rumelhart, D. E., & McClelland, J. L. (Eds.). (1986). *Parallel distributed processing: Vol. 1. Foundations*. Cambridge: MIT Press.

Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review*, **84**, 1-66.

Shiffrin, R. M. (1988). Attention. In R. C. Atkinson, R. J. Hernstein, G. Lindzey, & R. D. Luce (Eds.), *Steven's handbook of experimental psychology: Vol. 2. Learning and cognition* (2nd ed., pp. 739-811). New York: Wiley.

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. *Psychological Review*, **84**, 127-190.

Staller, S.J., Pelter, A.L., Brimacombe, J.A., Mecklenberg, D., & Arndt, P. (1991). Pediatric performance with the Nucleus 22-Channel Cochlear Implant System. *American Journal of Otology*, **12**, 126-136.

Sternberg, S. (1966). High-speed scanning in human memory. *Science*, **153**, 652-654.

Sternberg, S. (1969). Memory-scanning: Mental processes revealed by reaction-time experiments. *American Scientist*, **57**, 421-457.

Svirsky, M.A. (1996). Speech production and language development in pediatric cochlear implant users. Paper presented at the 131st meeting of the Acoustical Society of America, Indianapolis.

Waltzman, S.B., Cohen, N.L., Gomolin, R.H., Shapiro, W.H., Ozdaman, S.R. & Hoffman, R.A. (1994). Long-term results of early cochlear implantation in congenitally and prelingually deafened children. *American Journal of Otology*, **15**, 9-13.

Waltzman, S.B., Cohen, N.L., Gomolin, R.H., Green, Shapiro, W.H., Hoffman, R.A., & Roland. (1997). Open set speech perception in congenitally deaf children using cochlear implants. *American Journal of Otology*, **12**, 342-349.

Waugh, N. C., & Norman, D. A. (1965). Primary memory. *Psychological Review*, **72**, 89-104.

Zwolan, T.A., Zimmerman-Phillips, S., Asbaugh, C.J., Hieber, S.J, Kileny, P.R. & Telian, S.A. (1997). Cochlear implantation of children with minimal open-set speech recognition skills. *Ear & Hearing*, **18**, 240-251.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 21 (1996-1997)
Indiana University

Performance of Normal-Hearing Children on Open-Set Speech Perception Tests[1]

Melissa Kluck,[2] David B. Pisoni,[3] and Karen Iler Kirk[3]

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405

---

[1] This research was supported by NIH Research Grant DC-00111 to Indiana University in Bloomington, IN.
[2] Now at the Department of Communication Sciences and Disorders, Northwestern University, Chicago, IL
[3] Also DeVault Otologic Laboratory, Indiana University School of Medicine, Indianapolis, IN.

Performance of Normal-Hearing Children on Open-Set Speech Perception Tests

Abstract. The Phonetically Balanced Kindergarten Test (PBK), an open-set test of word recognition is typically included in test batteries designed to assess the speech perception skills of profoundly deaf children with cochlear implants. Many pediatric cochlear implant users have a great deal of difficulty with this test. Two new open set tests, the Lexical Neighborhood Test (LNT), and Multisyllabic Lexical Neighborhood Test (MLNT) (Kirk, Pisoni & Osberger, 1995), have been developed with the framework of the Neighborhood Activation Model (NAM) (Luce, 1986) of spoken word recognition. The LNT and MLNT are based on the lexical characteristics of word frequency and neighborhood density, and include words found in the vocabularies of children age three to five. Results from these tests with pediatric cochlear implant users have shown that their lexicons appear to be organized into similarity neighborhoods, and these neighborhoods are accessed in open-set word recognition tests. The present study investigates the speech perception abilities of normal hearing children ages three and four using the PBK, LNT and MLNT. The study was also designed to assess test-retest reliability for these tests using normal hearing children. Each child was first screened to ensure normal hearing, and measures of each child's vocabularies were obtained. Normal hearing three and four year old children performed extremely well, with scores near ceiling on all three tests. Because of the lack of variance in the scores, test-retest reliability could not be assessed using this population. The study did demonstrate, however, that it is reasonable to expect normal hearing three- and four-year old children to recognize all the words from these three open-set speech perception tests at very high levels of performance. These results can be used as a benchmark for children with hearing impairments who score poorly on the PBK, LNT and MLNT.

Introduction

This study is concerned with assessing the spoken word recognition abilities of children three to four years old. Several justifications exist for testing speech perception abilities of young children. First, these measures can aid in assessing rate of language development in normal as well as hearing or language impaired children, and are often used to detect and diagnose language delay. Second, a child's speech perception abilities may also provide insight about perceptual and cognitive skills important for language and reading. In the case of hearing impaired children, speech perception tests are critical in assessing progress in language skills, reading, and overall cognitive development, and as aids in planning education, rehabilitation, and speech therapy (Kirk, Diefendorf, Pisoni & Robbins, 1997).

Over the years, two types of word recognition tests have been used with children: closed-set tests and open-set tests. Closed-set tests require a child to choose between several possible responses, whereas in open-set tests the potential responses are unlimited. Closed-set tests are useful when testing perceptual abilities of children who cannot speak clearly or write, because the procedures force the subject to respond by pointing to pictures. Open-set tests are not appropriate for all children. Successful performance on open-set tests requires that the child be able to hear well enough to perceive the stimulus pattern, encode it, and then represent it in memory, and speak intelligibly enough so that the examiner can understand the responses when the child attempts to reproduce it at the time of test (Kirk et al., 1997).

Since the late 1940's, the Phonetically Balanced Kindergarten test (PBK) (Haskins, 1949), an open-set, monosyllabic speech perception test, has been used for testing young children who are profoundly deaf. More recently, this test has been used with children who have received cochlear implants (Kirk et al., 1997). The test consists of three lists, each with fifty words each, chosen from the International Kindergarten Union vocabulary lists. Each list is phonetically balanced. That is, the frequency of different phonemes reflects phoneme distribution in normal conversational speech.

With the exception of a small number of pediatric cochlear users who perform exceptionally well on this test, the so-called "Stars", the PBK is extremely difficult for many deaf children (Pisoni, Svirsky, Kirk & Miyamoto, 1997). Despite the difficulty of this test, it is routinely included in the standard battery of tests used to assess progress of cochlear implant users at Central Institute for the Deaf (CID), Indiana University School of Medicine, and many other cochlear implant centers around the world.

Although the PBK was originally designed for normal-hearing young children, and has been used by clinical audiologists for many years, it was not based on any theory or model of speech perception. Instead, it was developed primarily on empirical criteria. Low scores on this test by hearing-impaired children may be due to many factors. For example, the PBK test may be inappropriate for cochlear implant users because the vocabulary is too difficult (Osberger, Miyamoto, Zimmerman-Phillips, Kemick, Stroer, Firszt & Novak, 1991). Alternatively, perhaps deaf children have trouble perceiving isolated words in open-set tests where there are no response alternatives provided. To address these criticisms, specifically the issue about unfamiliar words, Kirk, Pisoni and Osberger (1995) recently developed several new open-set tests of speech perception to assess children using cochlear implants. These tests are firmly grounded in recent theoretical work on spoken word recognition (Luce, Pisoni & Goldinger, 1990). All the words were selected to be highly familiar to young children. The Lexical Neighborhood Test (LNT) contains two lists of 50 mono-syllabic words. The Multisyllabic Lexical Neighborhood Test (MLNT) contains two lists of 24 words, which are 2-3 syllables in length.

These new word recognition tests are theoretically motivated and make a number of specific predictions. That is, they were designed to test certain principles of speech perception and spoken word recognition processes used by normal hearing subjects. The tests were developed to provide descriptive data and theoretical insights into the underlying perceptual processes employed by pediatric cochlear implant users. The tests were also designed to be easier for cochlear implant users to carry out. Because the tests are easier, they should be more useful in discriminating between open-set test abilities of different populations of hearing-impaired children.

The LNT and MLNT tests are based on the Neighborhood Activation Model (NAM) (Luce, 1986) of spoken word recognition. The Neighborhood Activation Model assumes that two important characteristics of a word are its frequency (i.e., how often a word appears in the language), and its lexical neighborhood, (i.e., the number of words which are phonetically similar to it). These two factors have been shown to affect spoken word recognition when isolated words are presented to adults in noise (Luce, 1986; Goldinger, Luce, & Pisoni, 1989; Luce et al., 1990). A "lexical neighborhood" refers to the number of words which differ from a target word by only one phoneme. For example, the lexical neighbors of "bat" include the words *"cat"," ban"*, and *"bath"*. If a word has many lexical neighbors, it is assumed to reside in a "dense" lexical neighborhood. In contrast, if a word has few lexical neighbors, it is assumed to reside in a "sparse" lexical neighborhood.

Luce (1986) and Luce et al. (1990) have shown that normal hearing adult listeners find it easier to recognize high frequency words from sparse lexical neighborhoods (i.e., "easy" words) than low frequency words from dense lexical neighborhoods (i.e., "hard" words). The Neighborhood Activation Model has received a great deal of support from word recognition studies using adult listeners (Luce et al., 1990). The most important theoretical claim of the NAM model is that spoken words are recognized relationally in the context of other words in the listener's mental lexicon using a two-step process of acoustic-phonetic activation followed by lexical selection. Frequency and acoustic-phonetic similarity play independent roles in this overall process.

Lexical characteristics of words also have been shown to affect word recognition in adults with mild to moderate hearing loss (Pisoni, Miyamoto, Kirk, Sommers & Osberger, 1994). These findings encouraged the development of a test based on lexical characteristics of words to be used for testing children with profound hearing loss who are using cochlear implants. The LNT and MLNT tests were designed to measure speech perception skills and to provide new information about the way in which children with cochlear implants organize and access spoken words from memory. Differences in lexical neighborhood density and word frequency were used to generate lists of "easy" and "hard" words for these new tests, based on lexical characteristics of each word using assumptions derived from the NAM model (Kirk et al,. 1995).

The LNT contains two "easy" and two "hard" word lists. All words in this test are monosyllabic and exclude proper nouns, possessives, contractions, plurals and inflected forms. Words were chosen according to mean ratings of frequency and neighborhood density from Logan's (1992) earlier analysis of a large database of children's utterances. Logan used the CHILDES (Child Language Data Exchange System) database (MacWhinney & Snow, 1985), and computed lexical characteristics of words contained in the vocabularies of children aged one to five. Test items used for the LNT and MLNT were selected from words produced by children (ages three to five) contained in Logan's (1992) analyses. These ratings were calculated using the utterances analyzed by Logan. "Easy" words had frequencies above the median, and neighborhood densities below the median. The opposite is true for "hard" words. The MLNT contains two "easy" and two "hard" words, each with 12 words. These stimuli were chosen in a manner identical to that of all words included in the LNT, except the words are multi-syllabic.

Kirk et al. (1995) reported significant differences in lexical characteristics of the PBK, LNT and MLNT word lists, when rated for familiarity by normal hearing adult listeners. Although all three word lists were found to be highly familiar to adult listeners, important differences were found in neighborhood density. LNT "hard" words had the highest scores for neighborhood density, followed PBK words, and then LNT "easy" words. Using a computerized database, Kirk et al. (1995) investigated how familiar the PBK words were to young children, and found that only 31% of the PBK words were included in he CHILDES database analyzed by Logan (1992), which suggests that many of the PBK words are not common in the vocabularies of children under the age of five. This result suggests that the vocabulary used in the PBK may be unfamiliar to young children and pediatric cochlear implant users, which may be one of the reasons many deaf children with cochlear implants perform so poorly on this test. All of the LNT and MLNT words were chosen from words known to be in the vocabularies of children aged 1 to 5. Thus, these familiar words are more likely to be in the vocabularies of young children.

When pediatric cochlear implant users were assessed using the PBK, LNT and MLNT, Kirk et al. (1995) found significant differences in correct identification of the words on the three word lists. Word length and lexical characteristics had a significant effect of identification scores. Specifically, subjects' performance was highest on the MLNT, followed by LNT, and then PBK. Within the LNT and MLNT, scores on the "easy" lists were consistently higher than scores on the "hard" lists.

Kirk et al. (1995) concluded that the lexical characteristics of words on these tests affects word recognition scores of pediatric cochlear implant users. Based on the observed pattern of scores across the LNT and MLNT tests, Kirk et al. (1995) suggested that the lexicons of pediatric cochlear implant users are organized in a manner that is similar to normal hearing children and adults. That is, words are organized in long term memory according to similarity neighborhoods and frequency. Kirk et al. (1995) also found that longer words i.e. MLNT were easier to identify than short words (i.e., LNT), suggesting that these hearing-impaired children use the length of a word to discriminate and select spoken words from lexical memory.

Because no baseline data have been obtained yet for the LNT and MLNT using normal hearing children, we administered the PBK, LNT and MLNT to normal hearing three- and four-year old children. We also collected data on their vocabulary knowledge. The present study had two primary goals. The first goal was to examine test-retest reliability of the LNT and MLNT word lists. Test-retest reliability can be established by repeating the administration of an identical test after a reasonable amount of time has passed (Anastasi, 1961) to determine if the scores are repeatable. The second goal of the study was to provide normal hearing control data for the LNT and MLNT test. We expected to find effects of age on performance for these word recognition tests, such that older children should do better than younger subjects. We were also interested in knowing at what age is it reasonable to use the LNT and MLNT with children who have some hearing loss. We addressed these problems using two groups of young children who had normal hearing and no known language or other developmental delays.

Methods

Participants

All subjects were recruited from a database of children in the Bloomington, Indiana area. Letters and subsequent follow-up phone calls were used to inform parents of the experiment. Two groups of subjects were used. Eight four-year olds, 2 males and 6 females with an average age of 4 years 9 months participated in one group. One four-year old male did not return for the second session. Twenty-two three-year olds also participated in the other group: 8 males and 14 females, with an average age of 3 years 4 months. One three- year old female did not return for the second session. One three-year old male did not complete the first session, and one three-year old female did not complete either session. Only scores from completed tests were used in the final data analysis.

Test Materials

Three sets of words were used in constructing the audio recordings of the test stimuli. The first set was made up of the three PBK 50-word lists. The second set of words was made up of both LNT 50-word lists, each containing 25 "easy" and 25 "hard" words. The third set was made up of both MLNT 24-word lists, each containing 12 "easy" and 12 "hard" words. Thus, the total number of test stimuli was 298: 150 PBK, 100 LNT and 48 MLNT words.

The test words were read in isolation by a male talker. The stimuli were recorded in a single walled sound attenuated chamber (IAC No. 402) using a head mounted close talking microphone (Shure SM98A). The analog signal was low-pass filtered at 10 kHz (TDT FT5) and digitized directly to disk at 22 kHz using 16 bit resolution with the TDT System II interfaced to a 486 DX-66 based PC. Following the digital recordings, the waveform files were down-sampled to 20 kHz. To ensure uniform presentation levels, overall RMS (root mean squared) amplitudes were digitally equated for all the stimulus files.

Procedure

Subjects were screened on the first day of the study for normal hearing at 250, 500, 1,000, 2,000 and 4,000 Hz at 20 dB HL using a Maico Hearing Instruments audiometer (MA27) and TDH-39P headphones. Both ears were screened separately. Vocabulary scores were then collected using the revised version of the Peabody Picture Vocabulary Test (PPVT-R) (Dunn & Dunn, 1981).

The perceptual data for the three word recognition tests were collected in two separate sessions that were conducted between 2 and 9 days apart, with a mean time lapse of 5.0 days (SD = 2.152) for the three-year olds, and between 2 and 6 days apart with mean time lapse of 3.7 days (SD = 1.7) for the four-year olds. In each session, subjects heard one of three 50-word PBK lists, counterbalanced across subjects. Then, the LNT and MLNT lists were presented, in a counterbalanced order across subjects. Within the LNT and MLNT word sets, presentation order of lists was also counterbalanced. Stimuli were randomized within all lists. Subjects heard the same lists in the same order in the second session.

An IBM Thinkpad Computer (Model 750CS) was used to playback stimuli and record responses entered by the examiner. Stimuli were reproduced using a loudspeaker (Acoustics Research Instruments). The output level of the speech was set at 70 dB SPL, as determined by a hand-held sound pressure level meter (Triplett 370) that was placed at the approximate position of the subjects' head. Responses were transcribed immediately by the experimenter, using a computer program specifically designed to present stimuli, and record the examiner's transcription of the subject's response. A tape recorder was also used to record all subjects' responses during a test session.

During the first session, parents completed the Language Development Survey (Rescorla, 1991), and the MacArthur Communicative Development Inventory (MacArthur, 1993). These scales are used to detect language delay. During the second session, parents completed a familiarity task (FAM) of their child's vocabulary using the PBK, LNT, and MLNT word lists. This task required the parent to rate their child's familiarity with each of the test words on a scale from 1 to 7. Each parent was also asked to complete a survey of their own vocabulary knowledge (PFAM). Four hundred and fifty words were chosen from the Hoosier Mental Lexicon (Nusbaum, Pisoni & Davis, 1984), which contains familiarity ratings for 20,000 words chosen from the *Merriam-Webster Pocket Dictionary*. Based on the Hoosier Mental Lexicon, 150 words were rated as highly familiar, 150 had medium familiarity ratings, and 150 had low familiarity ratings. Parents were asked to rate their familiarity with these words on a scale from 1 to 7.

Children received stickers and candy to motivate them to continue during the experiment. At the end of the first session, each child received a small toy. Upon completion of the experiment, subjects were paid $10, and received a T-shirt with the laboratory logo, and a certificate of participation.

Results

First, we will report demographic data for the subjects. Next we will consider the mean scores for the subjects on the three word recognition tests. Following this we discuss the test-retest reliability pertaining to the word recognition scores. Finally, we will look at the familiarity measures completed by the parents. In each section, data are separated by age group. Scores for the three-year old group are followed by scores from the four-year old group for all measures reported below.

Demographic Data

Table 1 displays demographic data for both groups of subjects. Within the three-year old group, twenty-two subjects passed the hearing screening on day 1, and participated in the first session. The average age for this group was 40 months (SD = 3.7). Scores from the Peabody Picture Vocabulary Test (PPVT-R) had a mean of 44.5 (SD = 10.0). Within the three-year old group, chronological age and PPVT-R scores were correlated ($r = +0.63$, $p < .005$).

Within the four-year old group, eight subjects passed the hearing screening and participated in the first session. The average age of this groups was 57 months (SD = 1.2). The average of the PPVT-R scores, reported in terms of mental age, was 62.6 (SD = 13.3).

Table 1

Demographic Data for Both Age Groups

	Mean (mo.)	SD (mo.)	N	PPVT-R (mo.)	SD (mo.)
3 year olds	40	3.7	22	44.5	10.0
4 year olds	57	1.2	8	62.6	13.3

Word Recognition Tests

Table 2 displays the means for the perception tests for the three-year old group on both days. Means reported for the LNT and MLNT tests were averaged across lists 1 and 2, so for each test, scores are reported for "easy" and "hard" lists. On both days subjects performed highest on MLNT "easy" words, followed by MLNT "hard" words. The means for these tests were all above 90.5%. Means on the LNT "easy" and LNT "hard" tests ranged between 87.7% for LNT "hard" Day 2, and 89.2% for LNT "easy" Day 1. Means were lowest for the PBK test. The mean for Day 1 was 84.0%, and the mean for Day 2 was 86.8%.

Table 2

Results of Perception Tests for 3 Year Olds

	LNT "easy"	LNT "hard"	MLNT "easy"	MLNT "hard"	PBK
Mean % day 1	89.2%[b]	89.1%[b]	90.7%[a]	90.5%[a]	84.0%
(SD)	8.0	9.1	7.9	10.7	7.9
Mean % day 2	88.4%[b]	87.7%[b]	92.7%[b]	91.5%[b]	86.8%[a]
(SD)	9.3	6.6	8.5	7.6	11.3

Note. [a] N = 21 for these tests. [b] N = 20 for these tests.

Table 3 reports the means for the perception tests on Day 1 and Day 2 for the four-year old group. Means reported for the LNT and MLNT were averaged across lists 1 and 2, as they were for the three-year old group. The means ranged between 96.0% and 99.5%. For both days means were lowest on LNT "hard" words. The mean for Day 1 on this test was 95.7%, and the mean for Day 2 was 96.6 %. The second lowest mean was for the PBK test, with a mean of 96.0% for Day 1, and 96.9% on Day 2. On Day 1 the highest scores were on MLNT "easy" words, with a mean of 99.5%. On Day 2 the highest scores were on MLNT "hard" words, with a mean of 98.8%.

Table 3

Results of Perception Tests for 4 Year Olds

	LNT "easy"	LNT "hard"	MLNT "easy"	MLNT "hard"	PBK
Mean % day 1	97.5%	95.7%	99.5%	98.4%	96.0%
(SD)	2.3	1.7	1.5	3.1	4.8
Mean % day 2 [a]	98.3%	96.6%	97.6%	98.8%	96.9%
(SD)	1.8	3.2	3.3	2.0	2.0

Note. [a] N = 7 for these tests.

Figure 1 displays the means on the PBK, LNT and MLNT tests for both age groups. In this figure, LNT and MLNT scores were averaged across "easy" and "hard" lists, and across lists 1 and 2. The top panel shows the means from Day 1, with the three-year olds on the left and the four-year olds on the right. The bottom panel shows the comparison of means for Day 2. Overall, subjects had the lowest scores on the PBK, followed by the LNT, and then the MLNT. This was true for both the three- and four-year olds on both days. Figure 1 also shows that the four -year olds performed consistently better than the three- year olds, and that there was less variability in the scores from the four- year olds than in the scores from the three- year olds. This figure also shows that scores are close to the ceiling for both groups of subjects.

Insert Figure 1

Figure 2 shows the comparisons between the "easy" and "hard" lists for the LNT and MLNT. Data from the three- and four-year olds are shown separately. The top panel shows the means for Day 1, and the bottom panel shows the means for Day 2. In all but one case, the scores were slightly higher on the "easy" word lists than the "hard" word lists, but in all cases they were consistently close to ceiling levels of performance for both groups of subjects. The one exception was the MLNT test for the four- year old group on Day 2, where the mean was slightly higher on the "hard" words than the "easy" words.

Insert Figure 2

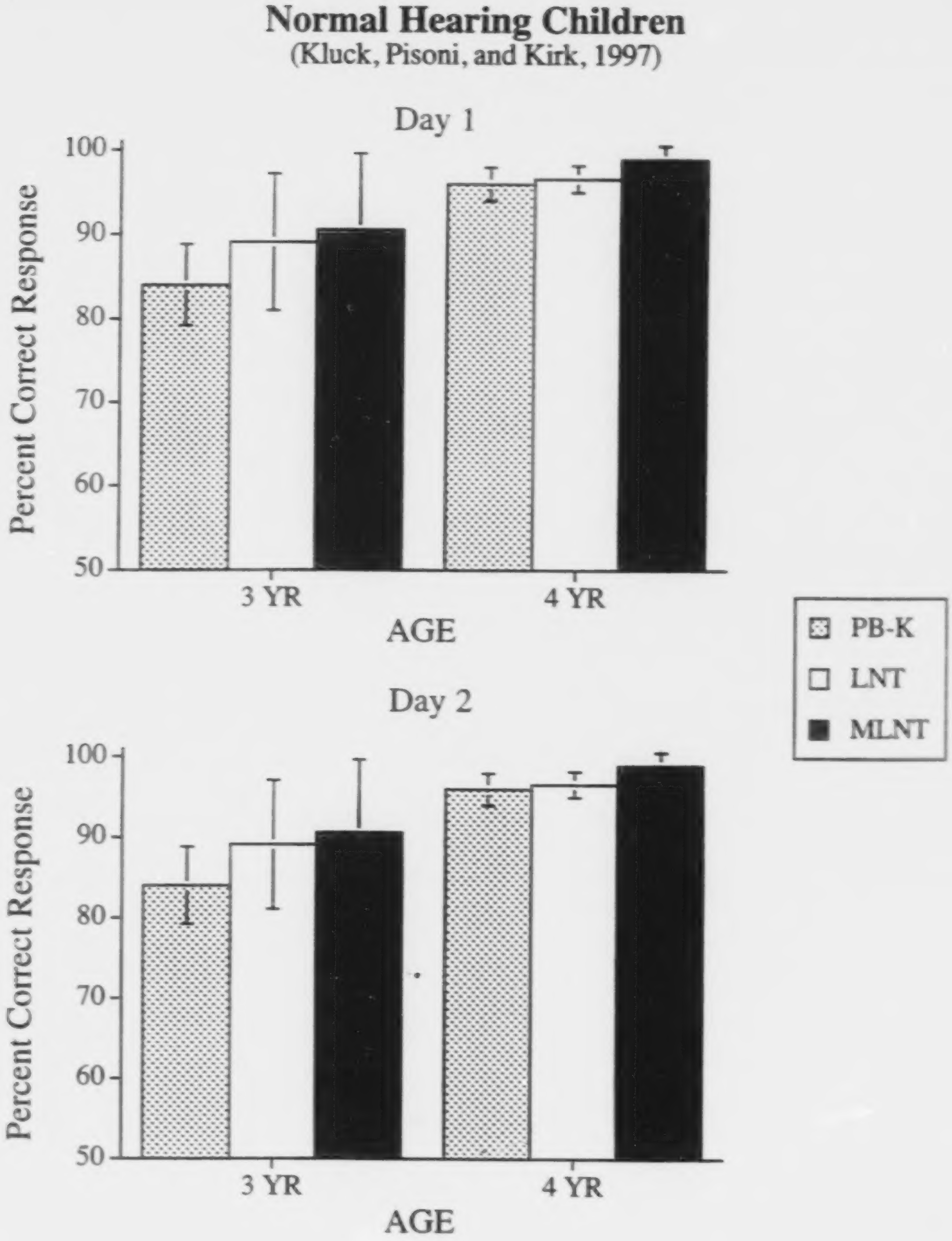
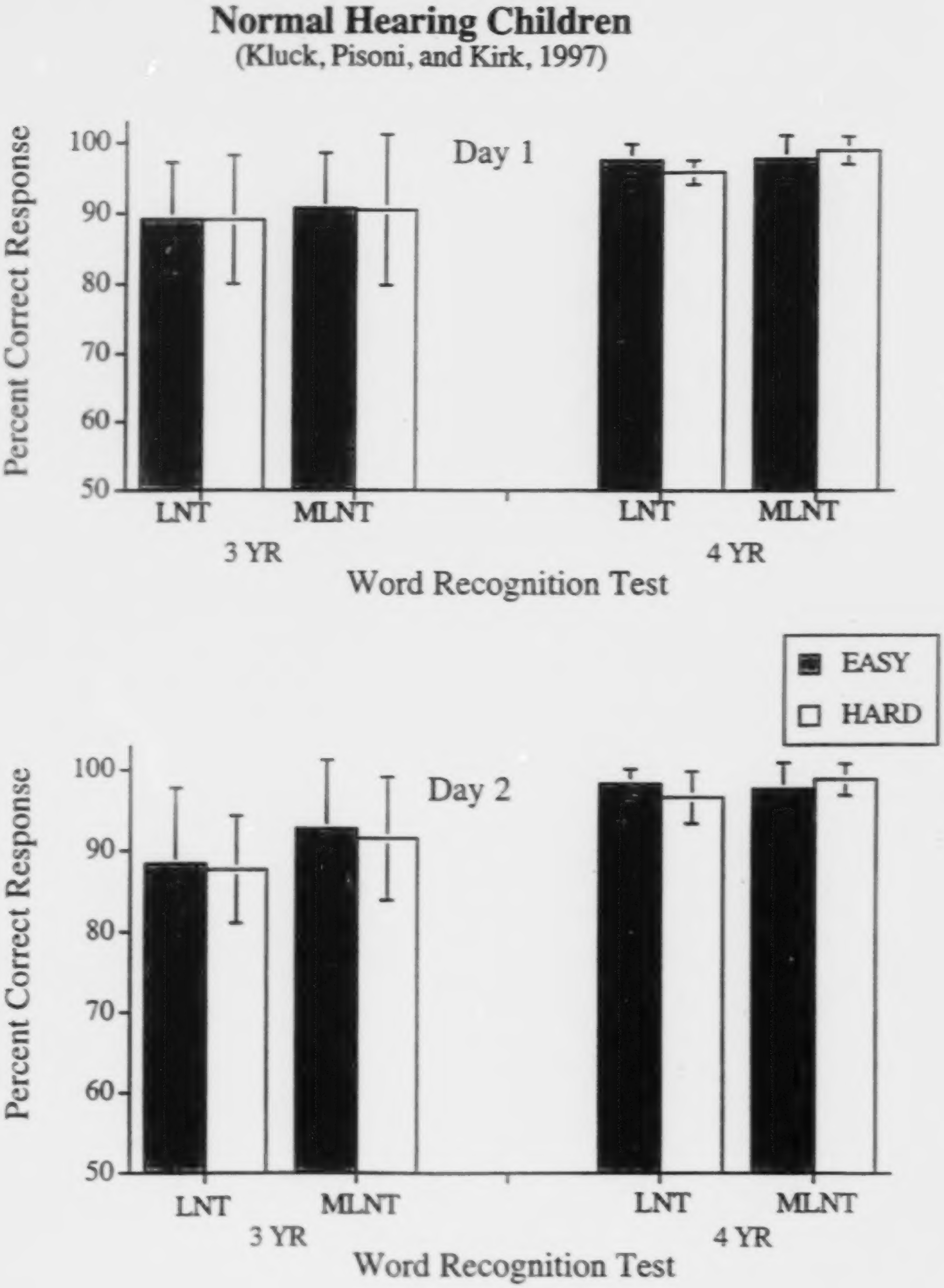


Figure 1. Percentage of words correctly identified for the LNT, MLNT and PBK, with scores broken down by age group, and day of testing.



Normal Hearing Children
(Kluck, Pisoni, and Kirk, 1997)

Figure 2. Comparisons between easy and hard word lists for the LNT and MLNT tests, broken down by age group and day of testing.

Paired t-tests were used to find differences between scores on the "easy" and "hard" word lists for each test. The "easy" and "hard" lists were significantly different for the LNT test on day 1 for the 4 year old group. There were no other significant differences between "easy" and "hard" word lists.

Test-retest Correlations

During this study, all subjects participated in two sessions in order to obtain data to assess test-retest reliability for all three perception tests. Because the scores were near ceiling on all of the perception tests, there was not sufficient variation or variability to report meaningful correlations. Figure 3 is a scatterplot of the PBK scores from both age groups, with Day 1 scores plotted using the x-axis, and Day 2 plotted using the y-axis. Figure 4 is a similar scatterplot using the scores from the LNT, and Figure 5 uses the scores from the MLNT. The purpose of these scatterplots is to demonstrate the high levels of performance, as well as the lack of variability in scores on these three tests.

Insert Figure 3

Insert Figure 4

Insert Figure 5

Parent Measures

The parents of the subjects completed several different measures during the course of the study. The Language Development Survey and MacArthur Communicative Development Inventory were used to detect language delay. All subjects knew almost every word on the vocabulary checklists of these measures, and showed adequate sentence formation ability. Therefore, there was no evidence of language delay in any of the participants.

Parents also completed familiarity ratings (FAM) of their child's knowledge of words used in the perception tests. For each word, parents used a rating scale from 1 to 7, with a high score indicating high familiarity with the word, and a low score indicating little or no knowledge of the word. The mean scores from both age groups on the three perception tests are reported in Table 4. The lowest scores in both age group were on the PBK FAM test, indicating that these words are judged by parents to be less familiar to the children than the LNT and MLNT words. The highest ratings for both age groups was on the LNT FAM test. Familiarity ratings for the four- year old group were consistently higher than the ratings for the three- year old group, although a two factor ANOVA showed that age did not have a significant effect on these familiarity ratings. This ANOVA did show an effect of word list ($F = 18.686$, $p<.0001$). Paired t-tests showed significant differences between all three word lists.

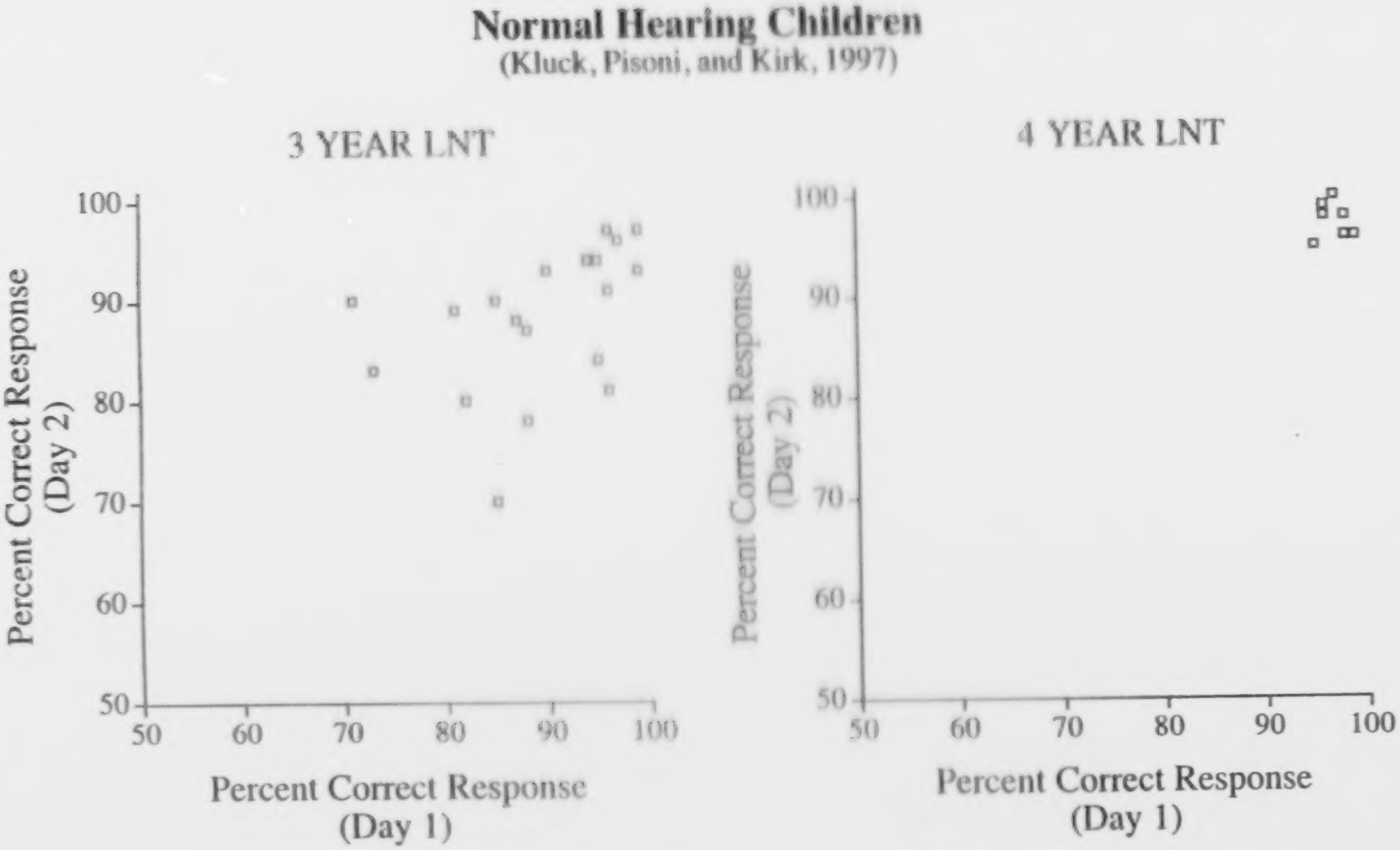


Figure 3. Scatterplots of percent correct scores for PBK for Day 1 and Day 2, with scores for the three-year olds on the left, and scores for the four-year olds on the right.

Normal Hearing Children
(Kluck, Pisoni, and Kirk, 1997)

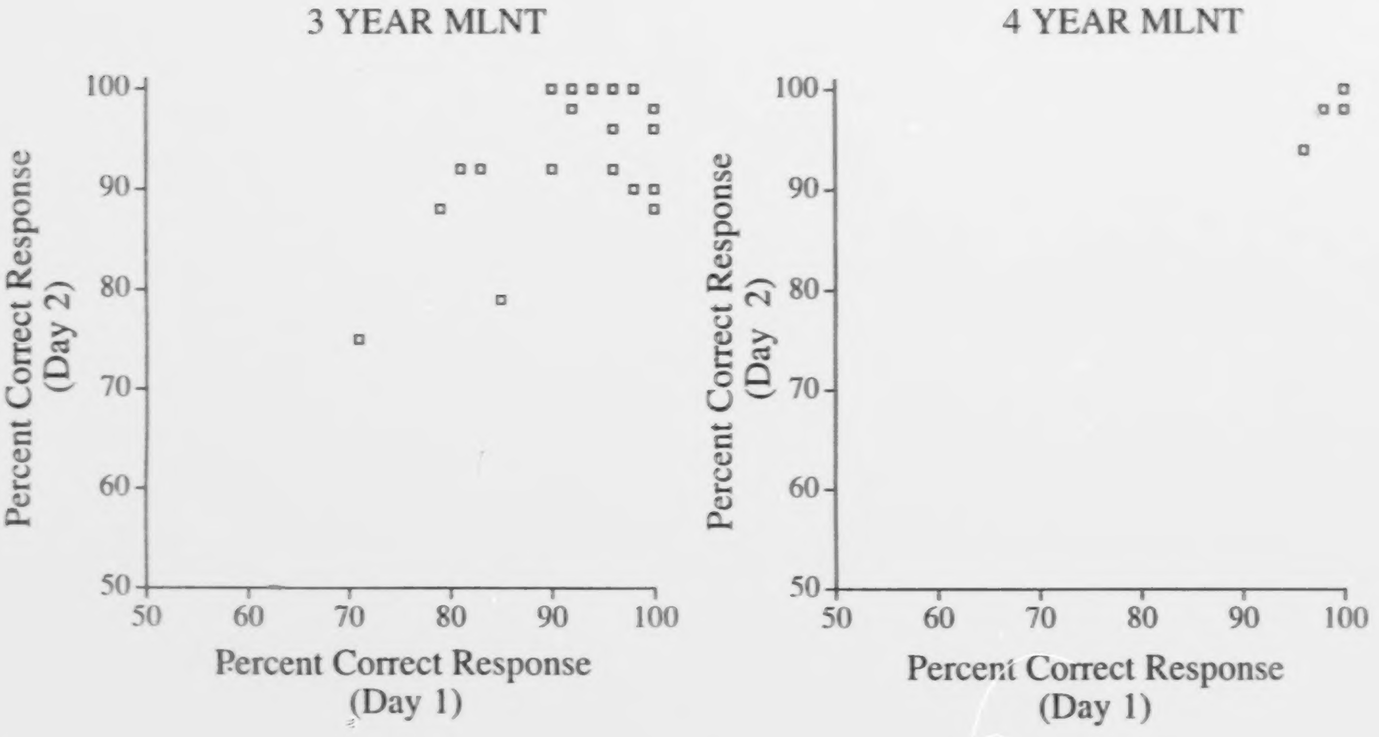


Figure 4. Scatterplots of percent correct scores for LNT for Day 1 and Day 2, with scores for the three-year olds on the left, and scores for the four-year olds on the right.

Normal Hearing Children

(Kluck, Pisoni, and Kirk, 1997)

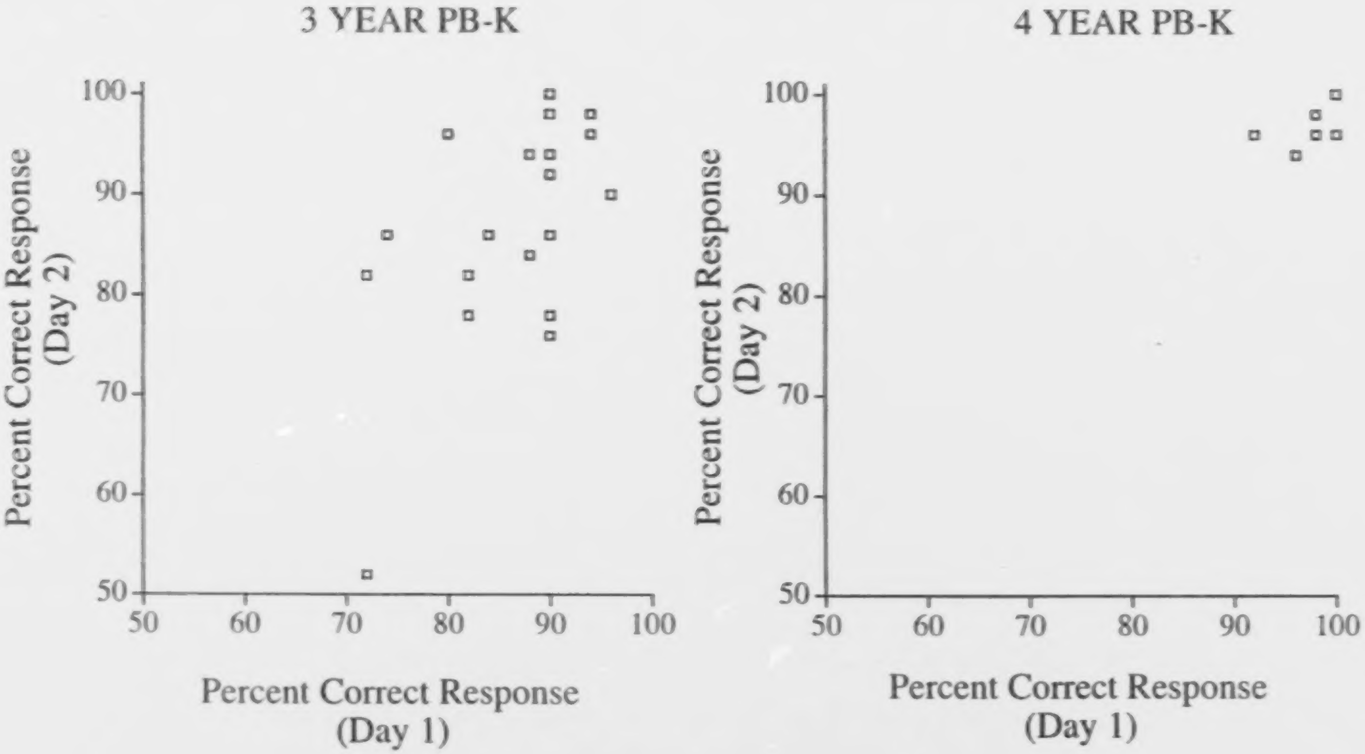


Figure 5. Scatterplots of percent correct scores for MLNT for Day 1 and Day 2, with scores for the three-year olds on the left, and scores for the four-year olds on the right.

Table 4

Parents familiarity ratings of their child's knowledge of words used in perception tests

	3 Year Olds [a]		4 Year Olds [b]	
	Mean (max. = 7)	(SD)	Mean (max. = 7)	(SD)
LNT FAM	6.74	.044	6.87	0.15
MLNT FAM	6.41	0.61	6.79	0.16
PBK FAM	5.95	1.0	6.58	0.37

Note. [a] N= 20 for these ratings. [b] N= 7 for these ratings.

The parents also completed a rating scale to assess their own vocabulary using the PFAM test. Four scores were obtained from the PFAM test: a low, medium, and high familiarity words, and an average of those three scores. Scores for both the three- and four- year olds are reported in Table 5.

Table 5

Means for PFAM test from both age groups

	PFAM HI	PFAM MED	PFAM LO	PFAM AVE
3 Year Olds	6.6	4.6	2.9	4.6
(SD) N=14	0.5	1.6	1.6	0.9
4 Year Olds	6.8	5.2	2.8	4.9
(SD) N=7	0.2	0.7	0.8	0.6

Discussion

Findings and Conclusions

The present investigation obtained several novel findings concerning the abilities of young children to recognize spoken words in isolation. The study was originally designed with two primary goals. First, we wanted to find the age at which a normal hearing child can successfully perform the LNT and MLNT tests. The results showed that normal hearing young children do very well on these three tasks, indicating that there may be another underlying reason, besides the difficulty of the vocabulary, that causes hearing impaired children to perform poorly on these three open-set tests. We found that three- and four- year old normal hearing children can successfully perform the PBK, LNT and MLNT open-set word recognition tests at near ceiling levels of performance. These findings can be used as a benchmark to compare performance abilities of young hearing-impaired children to normal hearing children.

In general, both groups performed at ceiling levels on the word recognition tests. Performance by both age groups was very high, and there was very little variance in the results. This is the main reason why we were unable to establish test-retest reliability for the PBK, LNT, and MLNT.

The construction and original design of the LNT and MLNT were motivated by several factors. These tests were intended to be easier open-set tests than the older PBK, in order to raise performance levels by cochlear implant users. These new tests were also theoretically based and were designed to study the effects of lexical neighborhood density and word frequency on word recognition. Results from this study did not provide any information about lexical organization in normal hearing 3 and 4 year old children, because subjects performed extremely well, at close to ceiling levels, on both "easy" and "hard" lists.

Familiarity ratings completed by parents provided additional information, in terms of numerical ratings, about their child's vocabulary knowledge. These measures are easily obtained, because parents fill out the rating while the child participates in the experiment.

Scores on the LNT FAM and MLNT FAM for the three- year olds were higher than scores on the PBK FAM. This finding supports the hypothesis from Kirk et al. (1995) that the vocabulary on the LNT and MLNT tests is simpler that the vocabulary on the PBK test.

In summary, the findings from this study can be used as a benchmark for the LNT and MLNT, where performance of hearing-impaired children on these tests can be compared to the results from the present study. We now know that the LNT and MLNT are appropriate for three- and four- year old children, and these findings can be used to compare the performance of a hearing-impaired child to the high performance of a normal hearing child.

Directions for Future Research

There are several important directions for future research using these tests. Further studies with normal hearing children may be able to provide more information about lexical organization in this population.

Several explanations are possible for the high performance of normal hearing children on open-set speech perception tests. One explanation may be that they know and understand the meaning of the target word. Another is that they are able imitate the sound sequence of the target word. To further investigate open-set test performance in normal hearing children, a future study could be conducted using words and pseudo-words which are unfamiliar to young children. Such a study would be informative because we would learn the reasons why young children are able to repeat the words they hear. For example, are they successful because they are familiar with the syllable structure of English, or do they actually need to know the meaning of the target word in order reproduce it correctly?

Future studies may also provide further evidence concerning the cause of poor performance of pediatric cochlear implant users on open-set tests. Does poor performance result because children with cochlear implants are unfamiliar with the vocabulary on the tests, or is there another more fundamental reason underlying their perceptual processes which are causing poor performance? Is it possible that the task of perceiving words in isolation, with no context, and then having to repeat isolated words from short term memory, is the cause of their low scores?

Finally, because the LNT and MLNT were designed to be used in testing pediatric cochlear implant users, perhaps it would be best to use this same population of subjects in establishing test-retest reliability for these new tests. It is most important that these tests are reliable when used for testing cochlear implant users, instead of with normal hearing children.

In summary, the present study of normal-hearing children has shown that this population can achieve high levels of performance on open-set word recognition tests. This finding, which may be used as an important bench mark to compare performance levels of various subject populations, encourages further investigation into the reasons behind the low levels of performance displayed by pediatric implant users. Study of normal hearing children may provide some insight into this problem. The results of the present study show that normal hearing three- year old children have little difficulty recognizing words from the PBK, LNT and MLNT tests.

The high scores obtained by normal hearing young children from the current study suggest that the poor performance of pediatric cochlear implant users may be due to other factors unrelated to the specific vocabulary items used on the open-set tests of word recognition. We believe it is important to try to identify some of these factors in future studies of children with cochlear implants.

References

Anastasi, A. *Psychological Testing 2nd Ed.* (1961). New York, NY: The Macmillan Company (p. 118).

Dunn, L. & Dunn, L. (1981). *Peabody Picture Vocabulary Test-Revised.* Circle Pines, MN: American Guidance Service.

Goldinger, S.D., Luce, P.A., & Pisoni, D.B., (1989). Priming lexical neighbors of spoken words: Effects of competition and inhibition. *Journal of Memory and Language, 28,* 501-518.

Haskins, H. (1949). *A phonetically balanced test of speech discrimination for children.* Unpublished master's thesis, Northwestern University, Evanston, IL.

Kirk, K.I., Pisoni, D.B., & Osberger, M.J. (1995). Lexical effects of spoken word recognition by pediatric cochlear implant users. *Ear & Hearing, 16,* 470-481.

Kirk, K.I., Diefendorf, A.O., Pisoni, D.B., & Robbins, A.M. (1997). Assessing speech perception in children. In Mendel, L.L., & Danhauer, J.L. (Ed.), *Audiologic Evaluation and Management and Speech Perception Assessment.* San Diego, CA: Singular Publishing Group, Inc.

Logan, J.S. (1992). A computational analysis of young children's lexicons. *Research on Spoken Language Processing Technical Report No. 8.* Bloomington, IN: Speech Research Laboratory, Indiana University.

Luce, P.A. (1986) Neighborhoods of words in the mental lexicon. *Research on Speech Perception Technical Report No. 6*. Bloomington, IN: Speech Research Laboratory, Indiana University.

Luce, P.A., Pisoni, D.B., & Goldinger, S.D., (1990). Similarity neighborhoods of spoken words. In G.T.M. Altmann (Ed.), *Cognitive Models of Speech Processing: Psycholinguistic and Computational Perspectives*. Cambridge, MA: MIT Press.

MacArthur Communicative Development Inventory: Words and Sentences. (1993). San Diego, CA: Singular Publishing Group, Inc.

MacWhinney, B., & Snow, C. (1985) The child language data exchange system. *Journal of Child Language, 12*, 271-296.

Nusbaum, H.C., Pisoni, D.B., Davis, C.K. (1984). Sizing Up the Hoosier Mental Lexicon: Measuring the Familiarity of 20,000 Words. *Research on Speech Perception Progress Report No. 10*. Bloomington, IN: Speech Research Laboratory, Indiana University.

Osberger, M.J., Miyamoto, R.T., Zimmerman-Phillips, S., Kemick, J.L., Stroer, B.S., Firszt, J.B., & Novak, M.A. (1991). Independent evaluation of the speech perception abilities of children with the Nucleus 22-channel cochlear implant system. *Ear and Hearing, 12*, 66S-80S.

Pisoni, D.B., Miyamoto, C., Kirk, K.I., Sommers, M.S., & Osberger, M.J. (1994). Sources of variability in speech perception by hearing-impaired listeners. Poster presented at the *17th Midwinter Research Meeting of the Association for Research in Otolaryngology*, St. Petersburg, FL., February 6-10, 1994.

Pisoni, D.B., Svirsky, M.A., Kirk, K.I., & Miyamoto, R.T. (1997). Looking at the "stars": A first report on the intercorrelations among measures of speech perception, intelligibility and language in pediatric cochlear implant users. Paper presented at the *5th International Cochlear Implant Conference*, New York, NY, May 1-3, 1997.

Rescorla, L. (1989). The language development survey: a screening tool for delayed language in toddlers. *American Speech-Language-Hearing Association, 54*, 587-599.

Effects of Talker, Rate and Amplitude Variation on Recognition Memory for Spoken Words[1]

Ann R. Bradlow[2], Lynne C. Nygaard[3] and David B. Pisoni

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405

---

[1] This research was supported by NIH-NIDCD Training Grant DC-00012 and NIH-NIDCD Research Grant DC-00111 to Indiana University. We are grateful to Luis Hernandez for technical support and to Thomas Palmeri for programming assistance. An earlier version of this study was presented at the 131st meeting of the Acoustical Society of America in Indianapolis, IN, May, 1996.

[2] Now at Department of Communication Sciences and Disorders, Northwestern University, Evanston, IL.

[3] Now at Department of Psychology, Emory University, Atlanta, GA.

A.R. Bradlow, L.C. Nygaard and D.B. Pisoni

Effects of Talker, Rate and Amplitude Variation on Recognition Memory for Spoken Words

Abstract. This study investigated the encoding of spoken words using a continuous recognition memory task. In Experiment 1, subjects judged whether each word in a list of spoken words was "old" (had occurred previously in the list) or "new." Subjects were more accurate at recognizing a word as "old" if it was repeated in the same voice, and at the same speaking rate; however, there was no recognition advantage for words repeated at the same overall amplitude. In Experiment 2, if subjects judged a word as "old" they were then required to provide an additional explicit judgment as to whether it was repeated in the same voice, rate, or amplitude. Subjects again showed an advantage in recognition memory for words repeated in the same voice and same speaking rate, but no advantage occurred for the amplitude condition. However, in all three conditions, subjects were able to detect whether an "old" word was repeated in the same voice, rate or amplitude. These data suggest that information about all three properties of spoken words is encoded and retained in memory and can be used in recognition tasks requiring explicit judgments.

Introduction

A growing body of research has begun to identify the effects of stimulus variability on a variety of speech perception and spoken word recognition tasks (e.g., Mullennix et al., 1989; Sommers et al., 1994). Other studies have also shown effects of stimulus variability on memory for spoken words (e.g., Martin et al., 1989; Goldinger et al., 1991; Palmeri et al., 1993; Nygaard et al., 1995; for reviews see Pisoni, 1993; 1997). These findings represent a novel approach to the long-standing issue of "perceptual constancy" in the face of a highly variable speech signal. Rather than actively seeking acoustic, articulatory or relational invariants that are supposed to guide the listener in accessing phoneme- and ultimately word-sized units (e.g., Joos, 1948; Ladefoged and Broadbent, 1957; Stevens and Blumstein, 1978; Kewley-Port, 1983; Halle, 1985; Nearey 1989; Johnson, 1990 and many others), this research directly investigates the effects of various sources of stimulus variability in the test materials. The general orientation of this research regards the inherent variability in the speech signal due to different talker- and other instance-specific characteristics as a useful source of information to the listener about the communicative situation (Laver, 1989; Laver and Trudgill, 1979), rather than a source of "noise" in the signal that is "stripped away" by the processes of speech perception and spoken word recognition (see Pisoni, 1997).

With respect to spoken word recognition, Mullennix et al. (1989) showed that word recognition accuracy decreased and response times increased when subjects were presented with lists of words produced by multiple talkers relative to a condition where subjects were presented with the identical words produced by only a single talker. Sommers et al. (1994) replicated this result with a different set of words and talkers. Additionally, in an attempt to understand the nature of the talker-variability effect found by Mullennix et al. (1989), Sommers et al. (1994) also investigated the effects of speaking-rate and overall amplitude variability on word recognition. The results of Sommers et al. (1994) replicated the findings reported by Mullennix et al. (1989). They showed a decrease in word identification scores for mixed-talker lists relative to single-talker lists. Furthermore, Sommers et al. (1994) showed a comparable decrease in word identification scores for mixed-rate lists relative to single-rate lists, but no decrease in word identification scores for mixed-amplitude lists relative to single-amplitude lists. These findings indicated that all sources of variability in the test materials do not produce similar effects on word recognition scores.

Sommers et al. (1994) suggested that the effects of talker and rate variability on word recognition may be due to the relevance of these dimensions for the perception of phonetic contrasts (Ladefoged and Broadbent, 1957; Miller, 1987). In contrast, variability in overall amplitude does not signal a phonetic contrast, and therefore variability along this dimension does not exert costly processing demands for word recognition.

With respect to memory for spoken words, Martin et al. (1989) found that subjects performed better in a serial recall task when the words within lists were produced by a single talker than when the words within each list were produced by multiple talkers. This difference in serial recall of spoken words was located in the primacy portion of the serial recall curve, that is, for the first three words in ten-word lists. Martin et al. (1989) proposed that this finding arose from the increased processing demands incurred by increased stimulus variability, and that these additional processing requirements interfered with subjects' abilities to maintain and rehearse information in working memory and to transfer this information to long-term memory.

Goldinger et al. (1991) investigated further the nature of talker variability effects on recall of spoken word lists by varying the rate of presentation of the items in the list to be recalled. Goldinger et al. (1991) hypothesized that rate of presentation would affect the subject's ability to encode the distinctive voice information for multiple-talker lists. If given enough rehearsal time, it was thought that subjects might be able to use the distinctive talker information as a retrieval cue, and thus the multiple-talker lists would be more accurately recalled than the single-talker lists. Indeed, Goldinger et al. (1991) found that at fast presentation rates (one word every 250 ms), words in the primacy portion of the single-talker lists were more accurately recalled than those from multiple-talker lists; whereas at slow presentation rates (one word every 4000 ms), this difference in recall accuracy was reversed. These results showed that information about a talker's voice is encoded and can be used as an effective retrieval cue under optimal conditions.

In a subsequent study, Nygaard et al. (1995) found that at fast presentation rates, items presented early in lists spoken either by a single talker or at a single speaking rate were better recalled than the same items spoken by multiple talkers or at multiple speaking rates, respectively. At a slow presentation rate, early items in the multiple-talker lists were better recalled than those in the single-talker lists; however, this reversal of recall accuracy was not obtained for the items in the multiple-rate lists relative to those in the single-rate lists. Rather, at the slow presentation rate, there was no difference between recall of items in the multiple- and single-rate lists. Furthermore, Nygaard et al. (1995) found no differences between serial recall of single- and multiple-amplitude lists at fast, as well as at slow presentation rates. Taken together, these results suggest that distinctive talker information is encoded in the long-term memory representation of spoken words, and if given sufficient rehearsal time, this additional distinctive information can be used as a retrieval cue by the listener. In contrast, the data from these serial recall experiments did not provide any evidence that either speaking rate or overall amplitude are encoded in long-term memory along with the linguistic content of a spoken word.

In a study of recognition memory for spoken words, Palmeri et al. (1993) found that detailed information about a talker's voice is retained in memory and facilitates recognition of a previously encountered word. Specifically, Palmeri et al. (1993) found that listeners were better at recognizing a word as a repeated item in a continuous list of spoken words when the word was repeated in the same voice that it was originally spoken in than when the voice differed from first to second repetition. Furthermore, Palmeri et al. (1993) showed that, when listeners recognized that the word was a repeated word in the list, they were also able to explicitly recognize whether the voice was the same or different as the first occurrence of the word.

401

Taken together, the findings of Goldinger et al. (1991), Nygaard et al. (1995) and Palmeri et al. (1993) have shown that specific talker characteristics can affect recall and recognition of spoken words. (See also Craik and Kirsner, 1974; Schacter and Church, 1992; Church and Schacter, 1994; Sheffert and Fowler (1995)). Furthermore, Nygaard et al. (1995) showed that variability in speaking rate can produce effects on the recall of spoken words, but that variability in overall amplitude does not. The purpose of the present study was to further investigate the role of different sources of variability in the encoding of spoken words in memory by comparing the effects of talker, rate and amplitude variability using a continuous recognition memory task. We hypothesized that a recognition memory task might be more sensitive in revealing the retention in long-term memory of stimulus dimensions such as speaking rate and overall amplitude than the serial recall task used by Nygaard et al. (1995) because a recognition task was thought to be less resource demanding than a recall task. Another goal of the present study was to provide additional data regarding the effects of different sources of variability on a variety of speech perception and word recognition tasks. Specifically, we wanted to know whether the distinct effects of talker, rate and amplitude variability on word identification found by Sommers et al. (1994) and by Nygaard et al. (1995) using serial recall tasks, would also be obtained in recognition memory. Thus, we hoped to be able to develop a more comprehensive understanding of the effects of different item-specific features on speech perception and spoken word recognition.

EXPERIMENT 1

Experiment 1 investigated whether subjects were more accurate at recognizing a word as "old" (i.e., had occurred previously in a list of spoken words) if it was repeated in the same voice (Condition 1), at the same speaking rate (Condition 2), and at the same amplitude (Condition 3). The voice condition was a replication of Palmeri et al. (1993); the rate and amplitude conditions were designed to extend the findings on voice to conditions where the stimuli incorporated other sources of variability.

Method

Subjects

One hundred and twenty students enrolled in undergraduate introductory psychology courses at Indiana University served as subjects. All subjects received partial course credit for their participation. All were native speakers of American English with no history of speech or hearing disorder at the time of testing.

Stimuli

The stimuli used in Experiment 1 came from a database of 200 words spoken by two talkers (one male and one female) at three different rates of speech (fast, medium, and slow). The words were selected from four 50-item phonetically balanced (PB) word lists (ANSI, 1971), and were originally recorded embedded in the carrier sentence, "Please say the word ______." For each rate of speech, the full set of 200 sentences was presented to the talkers in random order on a CRT screen located in a sound-attenuated booth (IAC 401A). Productions were monitored via a loudspeaker located outside the recording booth so that the mispronounced sentences could be noted and re-recorded. The stimuli were transduced with a Shure (SM98) microphone, and digitized on line in real-time via a 12-bit analog-to-digital converter (DT2801) at a sampling rate of 10 kHz. The stimuli were then low-pass filtered at 4.8kHz and the target words were digitally edited from the carrier sentences. The average root mean square amplitude of each of the stimuli was equated using a signal processing software package (Luce and Carrell, 1981). In order to create different presentation levels for the amplitude condition (Condition 3), high and low amplitude versions of the medium rate tokens from each of the two talkers were created. These tokens were generated by setting the maximum waveform amplitude level to a specified value. The remaining amplitude values in the digital files were then rescaled relative to this specified maximum. For the high and low amplitude sets, the maximum amplitude values were set at 60 dB SPL and 35 dB SPL, respectively. All other stimuli were leveled at 50 dB SPL.

For each of the three conditions (talker, rate, and amplitude) eight separate word lists were constructed in which each test word was presented and then repeated once after a lag of 2, 8, 16 or 32 intervening words. Each list began with 15 practice trials, which were used to familiarize the subjects with the test procedure. None of these 15 words was repeated in the experiment. The next 30 trials were used to establish a memory load and were not used in the final data analyses. The rest of the list consisted of 144 test word pairs, and 21 filler items which were not included in the analysis. The test pairs were distributed evenly across the four lags, with half of the repetitions at each lag in the same voice, rate or amplitude and half in a different voice, rate or amplitude as the original presentation of the test word. The total number of words in each list was 354.

For all three conditions, the lag between the first and second repetition of a word was manipulated as a within-subject variable (2, 8, 16 or 32 words). For the talker condition (Condition 1), only the medium rate tokens were used, and the voice of the talker for the second repetition of the target words was a within-subject variable (same vs. different voice). Forty-two subjects participated in Condition 1. For the rate condition (Condition 2), only the fast and slow rate tokens from both talkers were used. For this condition, Talker was a between-subjects variable, with half the subjects responding to tokens produced by the male talker (n=20) and half responding to tokens produced by the female talker (n=20). The speaking rate of the second repetition of the target words was a within-subject variable (same vs. different rate). Finally, for the amplitude condition (Condition 3), only the medium rate tokens from both talkers were used, and Talker was a between-subjects variable, with half the subjects responding to tokens produced by the male talker (n=19) and half responding to tokens produced by the female talker (n=19). The overall amplitude of the second repetition of the target words was a within-subject variable (same vs. different amplitude).

Procedure

Subjects were tested in groups of five or fewer in a quiet room used for speech perception experiments. The presentation of stimuli and collection of responses was controlled by a PDP-11/34 computer. Each digital stimulus was output using a 12-bit digital-to-analog converter and was low-pass filtered at 4.8 kHz. The stimuli were presented binaurally over matched and calibrated headphones (TDH-39) at a comfortable listening level. On each trial, subjects heard a spoken word and had up to five seconds to enter a response of "old" (i.e., the word had appeared previously in the list of spoken words) or "new" (i.e., the word was new to the list). Subjects entered their responses on appropriately labeled two-button response boxes. If no response was entered after five seconds, that trial was not recorded and the program proceeded to the next trial. No feedback was provided. The entire session of 354 trials lasted approximated 25-35 minutes.

Results and Discussion

Figure 1 shows the item recognition accuracies for the same-talker and different-talker repetitions (Figure 1a), same-rate and different-rate repetitions (Figure 1b), and same-amplitude and different-amplitude repetitions (Figure 1c) as a function of lag. For the Talker condition, a 2-factor ANOVA with Lag (2, 8, 16, 32) and Repetition (same-talker, different-talker) as factors showed significant main effects for both factors. Accuracy decreased with increasing lag (F(3,328)=24.518, p<.0001), and same-talker repetitions were recognized better overall than different-talker repetitions (F(1,328)=5.516, p<.0194). The two-way interaction was not significant. This result replicates the previous findings of Palmeri et al. (1993) that there is a same-voice advantage for recognizing a word as a repeated item without any explicit instructions to the subjects to attend to the talker's voice.

Insert Figure 1 about here

For the Rate condition, a 3-factor repeated measures ANOVA with Lag (2, 8, 16, 32), Repetition (same-rate, different-rate), and Talker (male, female) as factors showed significant main effects for Lag and Repetition but not for Talker (indicating no difference in recognition memory for words spoken by a male or a female talker). Accuracy decreased with increasing lag (F(3,152)=17.057, p<.0001) and same-rate repetitions were better recognized than different-rate repetitions (F(1,152)=39.895, p<.0001). There was no main effect of Talker (F(1,152)=.323, p=.5708) and none of the interactions were significant indicating that regardless of the talker, there were consistent and reliable effects of Lag and Repetition. This finding extends the same-voice advantage found by Palmeri et al. (1993) to a different item-specific characteristic of speech, and thus demonstrates that both talker and rate information are encoded in memory along with the symbolic/linguistic information about a spoken word.

For the Amplitude condition, a 3-factor repeated measures ANOVA with Lag (2, 8, 16, 32), Repetition (same-amplitude, different-amplitude), and Talker (male, female) as factors showed significant main effects for Lag and Talker. Accuracy decreased with increasing lag (F(3,144)=38.474, p<.0001), and accuracy was generally higher for the male talker than for the female talker (F(1,144)=4.319, p<.0395). However, there was no main effect of Repetition, and none of the interactions were significant. Thus, while recognition accuracy decreased with increasing lag, there was no difference in recognition accuracy between the same-amplitude and different-amplitude trials. Furthermore, this pattern of results was obtained for both talkers even though the overall accuracy scores for the male talker were slightly higher than for the female talker (91.2% and 88.9% correct item recognition, respectively). The fact that there was no same-amplitude advantage for both talkers suggests that overall amplitude information may not be a property of speech that is encoded into long-term memory in the same way as talker and rate information, and that different item-specific stimulus characteristics can have distinct effects on speech perception and spoken word recognition.

In order to compare the overall level of discrimination between "old" and "new" items across the three conditions, we computed d' scores for each subject in each condition. The mean d' score in all three conditions was significantly greater than zero (p<.0001 in all three conditions by a one-sample t-test), indicating good discrimination in all conditions (see Table I). Furthermore, a one-factor ANOVA with Condition as the factor, showed a significant main effect of Condition (F(2,117)=5.198, p<.007). Post-hoc comparisons (Fisher's PLSD) showed a significant difference in d' for the Talker and Rate conditions (p<.002), and for the Rate and Amplitude conditions (p<.039). However, there was no difference in d' for the Talker and Amplitude conditions.

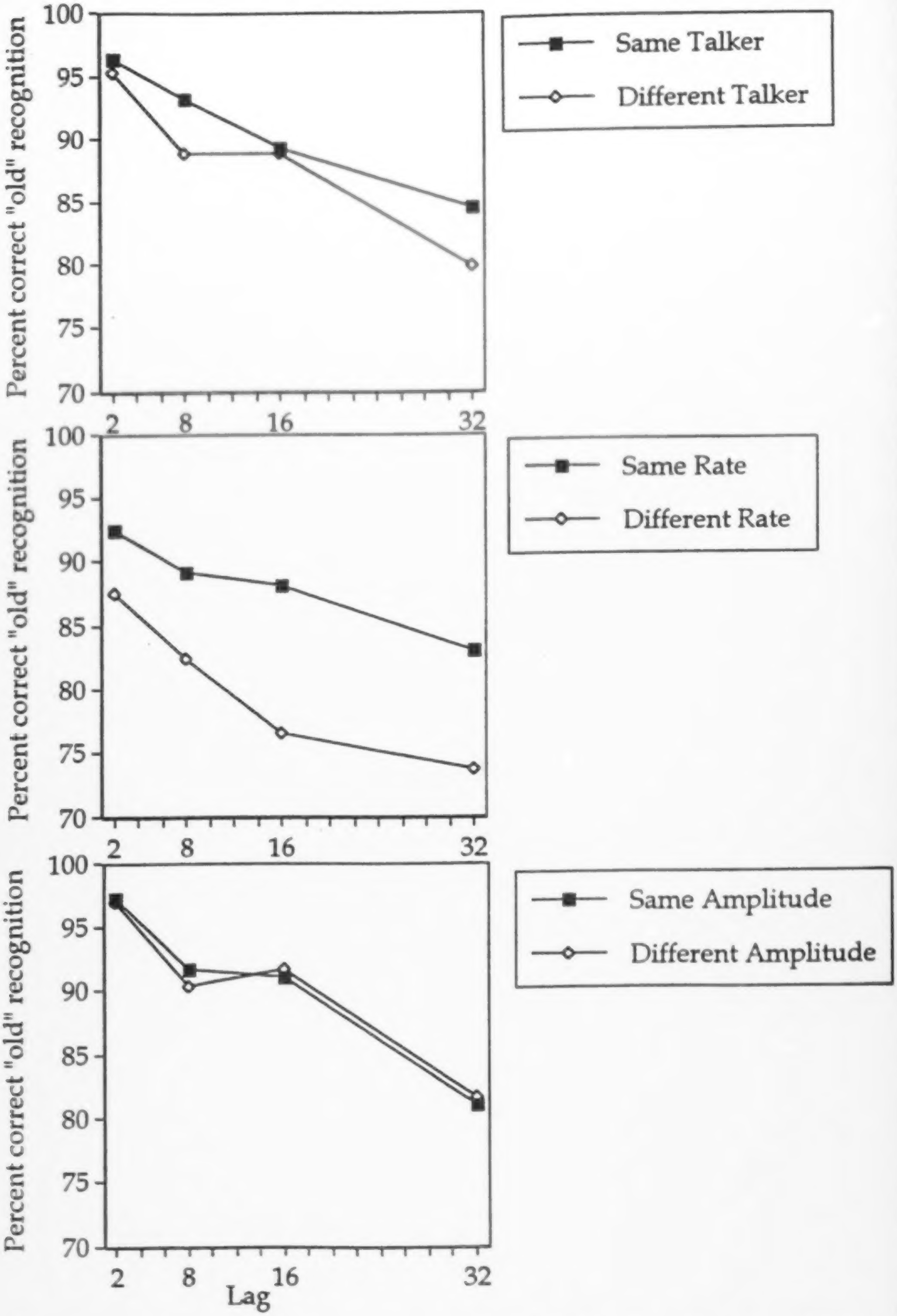


Figure 1. Item recognition accuracy scores as a function of lag from Experiment 1 for (a) the Talker condition, (b) the Rate condition, and (c) the Amplitude condition.

Table I

False Alarm Rates for Experiment 1.

Condition	Hit Rate	False Alarm Rate	d prime
Talker	81.0%	11.6%	2.17
Rate	80.3%	16.4%	1.91
Amplitude	81.5	13.5%	2.08

In summary, the results of Experiment 1 demonstrate that same talker and same rate trials were recognized better than different talker and different rate trials, respectively. In contrast, there was no difference in recognition memory for same and different amplitude trials. Thus, information about the talker's voice and speaking rate are encoded in the long-term memory representation of spoken words. However, there was no evidence that information about the overall amplitude of a spoken word is encoded in memory. The possibility remains, of course, that overall amplitude information may be retained in memory, but that when subjects are instructed to recognize the item as "new" or "old" they are unable to use this information as an implicit retrieval cue in this task. In order to evaluate this alternative another experiment was carried out.

EXPERIMENT 2

Experiment 2 was designed to investigate whether listeners can explicitly recognize changes in talker, rate and amplitude for a repeated word. Whereas in Experiment 1, subjects were not required to pay explicit attention to the voice, rate or amplitude of the test item, in Experiment 2, subjects were required to make an explicit judgment regarding a change in voice, rate or amplitude. We hypothesized that this would be a more sensitive test of the extent to which detailed information about the instance-specific characteristics of a spoken word are encoded in long-term memory. Specifically, we were interested in investigating the possibility that subjects are able to detect changes in overall amplitude even though a change in overall amplitude did not produce changes in recognition accuracy scores for words in the amplitude condition of Experiment 1.

Method

Subjects

One hundred and nineteen students enrolled in undergraduate introductory psychology courses at Indiana University served as subjects. All subjects received partial course credit for their participation. All were native speakers of American English and reported no history of speech or hearing disorder at the time of testing.

Stimuli and Procedure

The stimulus materials for Experiment 2 were identical to those used in Experiment 1. All aspects of the stimulus presentation and test conditions were identical to Experiment 1 except that in this experiment subjects were given three response categories rather than two. In Experiment 2, after hearing the spoken word, subjects had 5 seconds to identify the word as "new" if it had not occurred in the list

406

before, as "old-same" if it had occurred before and was repeated with the same voice (Condition 1), rate (Condition 2) or amplitude (Condition 3), or as "old-different" if it was repeated with a different voice (Condition 1), rate (Condition 2) or amplitude (Condition 3). Thus, in Experiment 2, in addition to recognizing a word as "old" or "new," subjects were also required to make an explicit judgment for the items recognized as "old" regarding voice, rate or amplitude variation from the first to second repetition of the word. A group of 33 subjects participated in the talker condition. For the rate condition, a group of 21 subjects was tested on stimuli spoken by the male talker, and a separate group of 21 subjects was tested on stimuli spoken by the female talker. For the amplitude condition, a separate group of 22 subjects was tested on each of the two stimulus sets (one from the male talker, one from the female talker).

Results and Discussion

Figure 2 shows the overall percentage of correct "old" item recognition for the talker condition (Figure 2a), the rate condition (Figure 2b) and the amplitude condition (Figure 2c). The accuracy scores shown in this figure represent all cases of correct "old" item recognition regardless of accuracy on the "same-different" judgment. This analysis allowed us to compare the pattern of results on the item recognition task across Experiments 1 and 2.

Insert Figure 2 about here

As shown in Figure 2, same talker trials were recognized better than different talker trials, same rate trials were recognized better than different rate trials, but there was no difference in recognition accuracy for same and different amplitude trials. This pattern of results is consistent with the results of Experiment 1. For the talker condition, a 2-factor ANOVA with Repetition (same talker or different talker) and Lag (2, 8, 16. 32) as factors showed main effects of both factors. Same-talker trials were better recognized than different-talker trials (F(1,256)=4.541, p=.0340), and recognition accuracy decreased with increasing lags (F(3,256)=13.258, p<.0001). The two-way interaction was not significant.

For the rate condition, a 3-factor repeated measures ANOVA with Repetition (same rate or different rate), Lag (2, 8, 16. 32), and Talker (male, female) as factors showed main effects of all three factors. Same-rate trials were better recognized than different-rate trials (F(1,160)=26.973, p<.0001), recognition accuracy decreased with increasing lags (F(3,160)=20.906, p<.0001), and recognition accuracy was slightly better for tokens produced by the male talker than for those produced by the female talker (mean difference = 2.94%, F(1,160)=4.815, p=.0297). None of the interactions involving Talker as a factor was significant indicating that the pattern of decreasing recognition accuracy with increasing lags, and across same-rate and different-rate trials, was consistent across both talkers. Similarly, the two-way interaction between Repetition and Lag was not significant.

For the amplitude condition, a 3-factor repeated measures ANOVA with Repetition (same amplitude or different amplitude), Lag (2, 8, 16. 32), and Talker (male, female) as factors showed a main effect of Lag, but no main effects of Repetition or Talker. None of the interactions was significant. As expected from the results of Experiment 1, recognition accuracy decreased with increasing lags (F(3,168)=48.820, p<.0001), but there was no same-amplitude advantage relative to different-amplitude trials. This pattern of results replicates the main results of Experiment 1 by providing evidence that information regarding the talker's voice and rate of speech are encoded in long-term memory along with the

407

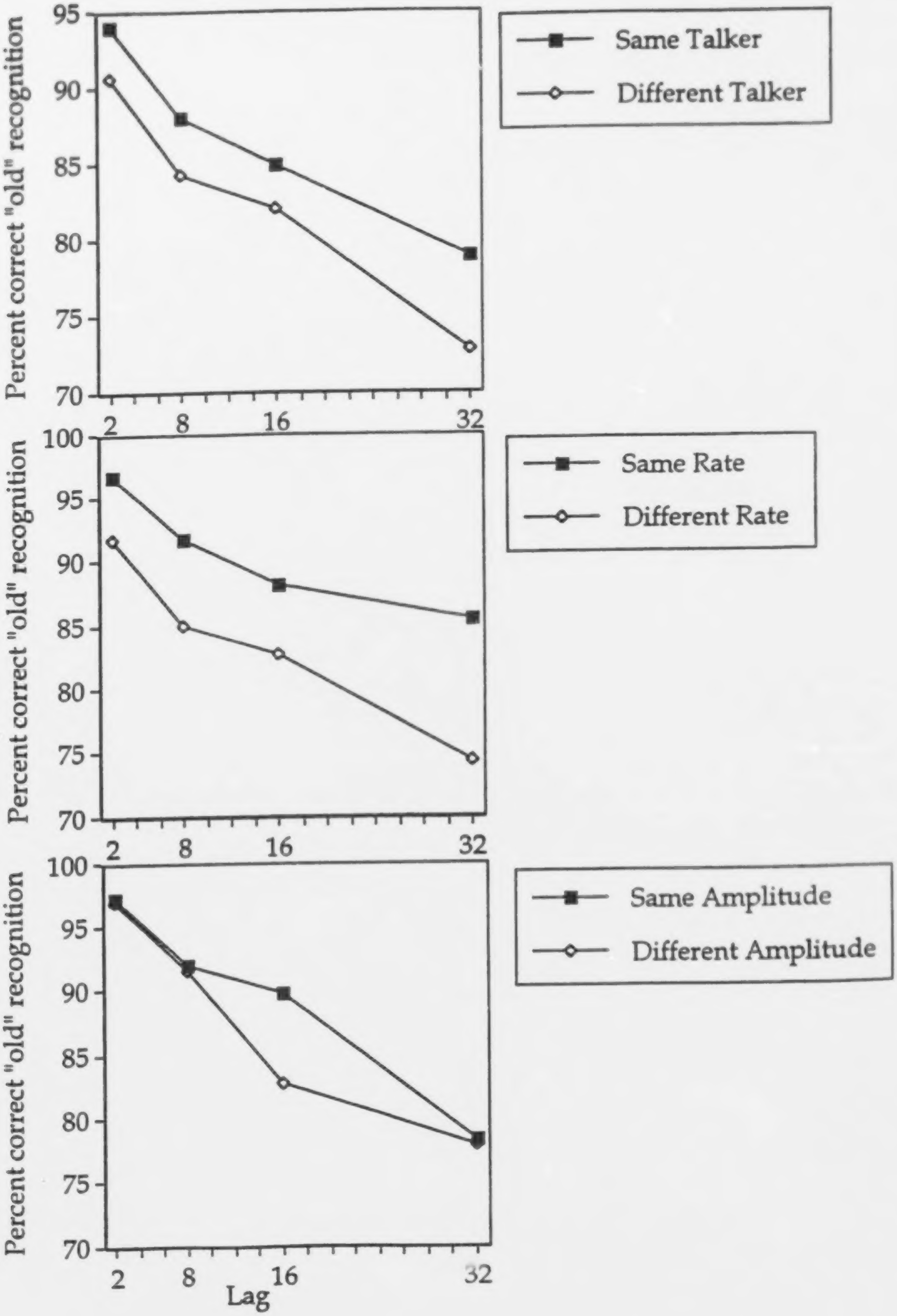


Figure 2. Item recognition accuracy scores as a function of lag from Experiment 2 for (a) the Talker condition, (b) the Rate condition, and (c) the Amplitude condition.

linguistic information about a spoken word. In contrast, once again, there was no evidence that information about overall amplitude is retained in long-term memory.

The similarity between the patterns of item recognition accuracy scores for the two experiments indicates that the additional response category for Experiment 2 did not alter the main effects of Lag and Repetition on item recognition accuracy. In order to assess directly the effect of the additional response category, separate repeated measures ANOVAs for each of the three conditions with Experiment (1 or 2) as the repeated measure were performed. For the Talker condition, the analysis showed the expected main effects of Lag ($F(3,256)=33.364$, $p<.0001$) and Repetition ($F(1,256)=8.552$, $p=.0038$). The two-way interaction between Lag and Repetition was not significant. There was also a significant main effect of Experiment ($F(1,256)=12.059$, $p=.0006$) due to generally higher accuracies for Experiment 1 than for Experiment 2 (means = 88.55% and 84.36%, respectively). None of the interactions involving the Experiment factor were significant indicating that the patterns of decreasing accuracy with increasing lag, and of higher accuracy for same-voice repetitions, were consistent across both experiments. For the Rate condition, there were main effects of Lag ($F(3,344)=40.025$, $p<.0001$) and Repetition ($F(1,344)=73.220$, $p<.0001$), but there was no effect of Experiment and none of the interactions were significant. Finally, for the amplitude condition, the main effect of Lag was significant ($F(3,304)=76.150$, $p<.0001$), and the main effect of Experiment was significant ($F(1,304)=7.398$, $p=.0069$) but there was no main effect of Repetition. As for the Talker condition, the effect of Experiment for the Amplitude condition was due to generally higher accuracies for Experiment 1 than for Experiment 2 (means = 90.21% and 87.82%, respectively). Thus, the additional response category in Experiment 2 resulted in slightly lower overall recognition accuracy scores for the Talker and Amplitude conditions. However, across all three conditions, the general pattern of results for the two experiments was consistent in showing a same-voice and same-rate advantage relative to different-voice and different-rate trials, respectively. Similarly, both experiments showed no same-amplitude advantage relative to different-amplitude trials.

In order to determine whether subjects can explicitly recognize variation in talker, rate and amplitude for items that were correctly identified as "old," d' scores were calculated for each condition at each lag. In this analysis, a "Hit" was defined as a response of "old/same" to a stimulus that was repeated with the same voice, rate or amplitude. A "False Alarm" was defined as a response of "old/same" to a stimulus that was repeated with different voice, rate or amplitude. Using this measure, we were able to determine if listeners can discriminate changes in talker, rate and amplitude, and thus establish whether detailed information along each of these dimensions was retained in memory.

Insert Figure 3 about here

Figure 3 shows the d' scores for all three conditions as a function of lag. Two main findings are shown in this figure. First, for all three conditions at all lags, the d' scores differed significantly from zero indicating that subjects were able to discriminate "old/same" from "old/different" trials in all cases. One sample t-tests for each condition at each lag confirmed that these d' scores were all significantly different from zero at the $p<.0001$ level. This finding suggests that, regardless of whether the instance-specific information affected recognition memory accuracy in the "old-new" task, listeners do retain highly detailed information in memory to the extent that variability along each of the three dimensions was explicitly detected. Second, variability along each of the three dimensions was detected with a different degree of accuracy: talker variability was detected better than rate variability which was detected better than amplitude variability. A two factor ANOVA with Condition (talker, rate, amplitude) and Lag (2, 8, 16, 32)

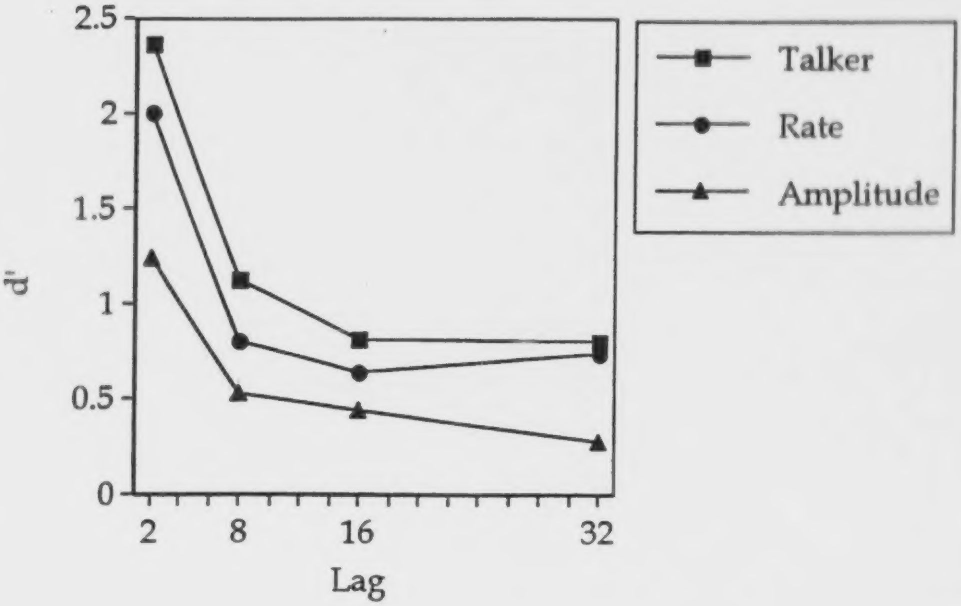


Figure 3. d' scores for all three conditions of Experiment 2 as a function of lag.

as factors showed main effect of both factors (Condition: F(2,476)=45.459, p<.0001; Lag: F(3,476)=110.988, p<.0001). The two-way interaction was also significant (F(6,476)=3.264, p=.0037). This finding suggests that, although fine details of the stimulus dimensions are retained in memory, certain dimensions represent more perceptually salient characteristics than others, and thus may produce more substantial effects on speech perception and spoken word recognition performance in different tasks.

The main goal of Experiment 2 was to investigate whether subjects were able to explicitly discriminate changes in talker, rate or amplitude for items that they recognized as repeated items (i.e., "old" items). In particular, we were interested in the results of this task for the amplitude condition where inconsistent amplitude information did not affect recognition memory performance. The results showed that subjects were indeed able to explicitly detect changes in talker, rate and amplitude. Thus, this task provided evidence that, even though all sources of variability do not function identically with respect to spoken word recognition, detailed information about the instance-specific characteristics of a spoken word is retained in memory along with the more abstract linguistic content of the word. These highly detailed memory representations even include information along an apparently linguistically irrelevant dimension such as overall amplitude.

GENERAL DISCUSSION

The overall goal of this study was to investigate the extent to which the neural representation of spoken words encodes detailed, instance-specific information. The results that emerged from this study complement the findings of earlier studies that have investigated the effects of talker, rate and amplitude variability on speech perception and memory for spoken words. The general pattern of results that has emerged from this set of experiments (summarized in Table II) suggests that information about all sources of variability is retained in long-term memory. However, the processing costs incurred by trial-to-trial variability along different stimulus dimensions varies for different properties of the speech signal.

A comparison of the effects of talker, rate, and amplitude variability on the tasks listed in Table II reveals a hierarchy in which amplitude, rate and talker variability have increasingly profound effects on speech perception and memory for spoken words. The relatively weak effect of amplitude variability is seen by the fact that experiments using all three tasks (word identification, serial recall, and continuous recognition) failed to find an effect of trial-to-trial changes in overall amplitude. In fact, the only evidence that overall amplitude information is retained in long-term memory comes from the task in which subjects were asked to *explicitly* identify variability along this dimension (present study, Experiment 2). In contrast, the stronger effect of rate variability was evident in all three tasks, where trial-to-trial changes in speaking rate resulted in decreased performance relative to trials with no change in speaking rate. For instance, word lists in which each word was spoken at a constant speaking rate were better identified when embedded in noise than identical lists spoken with multiple speaking rates (Sommers et al., 1994). Similarly, single-rate word lists were more accurately recalled than multiple-rate lists (Nygaard et al., 1995); and, consistent rate trials were better recognized in a continuous recognition memory task than trials in which the rate changed (present study, Experiment 2). The effects of talker variability are comparable to the effects of rate variability, however, a difference between the effects of talker and rate variability emerged in the serial recall task with long ISI's (Nygaard et al., 1995). When given enough time, the talker's voice was apparently encoded by the listeners in the long-term memory representation of the spoken words, and thus served as an identifying feature of the words. In this manner, the talker's voice functioned as a retrieval cue and aided the listener in the serial recall task to the extent that multiple talker lists were *better* recalled than single talker lists. In contrast, at long ISI's, the detrimental effect of multiple speaking rates was diminished only to the extent that multiple rate lists were recalled as well as single rate lists. Thus, the results of these

411

studies lead us to postulate a hierarchy of effects of stimulus variability on speech perception and memory for spoken words with talker variability having the most pervasive effects, rate variability having intermediate effects, and amplitude variability having the weakest effects.

Table II.

Summary of findings regarding the impact of talker, rate and amplitude variability on speech perception and memory for spoken words.

Source of Variability	Word Identification	Serial Recall		Recognition Memory	
		Short ISI's	Long ISI's	Item Recognition	Attribute Recognition
Talker	Single>Multiple[1,2]	Single>Multiple[3,4,5]	Multiple>Single[4,5]	Same>Different[6,7]	Yes[6,7]
Rate	Single>Multiple[2]	Single>Multiple[5]	Multiple=Single[5]	Same>Different[7]	Yes[7]
Amplitude	Single=Multiple[2]	Single=Multiple[5]	Multiple=Single[5]	Same=Different[7]	Yes[7]

[1] Mullennix et al. 1988
[2] Sommers et al. 1994
[3] Martin et al. 1989
[4] Goldinger et al. 1991
[5] Nygaard et al. 1995
[6] Palmeri et al. 1993
[7] Present study

At this point we can speculate as to the mechanism that underlies these different effects for different sources of variability. It is possible that these differences in the effects of talker, rate, and amplitude variability reflect differences in the complexity of the acoustic correlates of changes along these dimensions. In all of the experiments listed in Table II that investigated the effects of amplitude variability, a change in amplitude was achieved by simply setting the maximum level for each waveform to a specified value and then rescaling the remaining amplitude levels relative to that maximum. Thus, amplitude variability was a constant, uni-dimensional adjustment. In contrast, rate variability was more naturally achieved, and was thus variable and multi-dimensional in its acoustic correlates. Rate variability within a given speaker is not achieved by a constant "stretching" or "shrinking" of the acoustic waveform. Rather, certain acoustic segments are more dramatically reduced in duration than others when overall speaking rate is increased, and various other acoustic-phonetic changes (e.g., vowel reduction) occur in response to changes in speaking rate (e.g., Lehiste, 1972; Klatt, 1973, 1976; Port, 1981; Picheny et al., 1986, 1989; Uchanski et al., 1996). Thus, an increase or decrease in speaking rate is clearly a dynamic, multi-dimensional transformation of the speech signal. Similarly, a change in talker leads to a wide variety of acoustic-phonetic changes. Not only do different talkers differ in vocal tract shape and size, which leads to different spectro-temporal characteristics, but different talkers also differ in articulatory "style" (including speaking rate, dialect, and other idiosyncratic differences) which can lead to large differences in the acoustic waveform of a given word across various talkers (e.g., Fant, 1973; Joos, 1948; Peterson and Barney, 1952).

Thus, the extent of the effects of variability in talker, rate, and amplitude investigated by the experiments listed in Table II appear to be directly related to the complexity of the acoustic correlates that result from these sources of variability. From the listener's point of view then, it is possible that the simpler the acoustic transformation related to a given source of variability, the fewer the processing resources required to compensate for that variability, and consequently the less the impact of this variability on speech perception and memory for spoken words.

Another explanation for the differential effects of the different sources of variability on speech perception and memory for spoken words takes into account the relevance of each source of variability for the perception of phonetic contrasts. Variability in talker characteristics has been shown to have a significant impact on phonetic contrast perception. For example, Ladefoged and Broadbent (1957) found that vowel identification could be altered depending on the perceived talker characteristics of a precursor phrase, and Johnson (1990) showed that perceived speaker identity plays an important role in the F0 normalization of vowels. Similarly, several studies have demonstrated the rate dependency of phonetic processing for both vowels and consonants (e.g., Port, 1981; Summerfield, 1981; Miller, 1987; Miller and Volaitis, 1989). In contrast, overall amplitude variability does not, by itself, signal phonetic contrasts, and there does not appear to be an amplitude-dependency in speech perception that is comparable to talker- and rate-dependent phonetic processing. Thus, it is possible that the observed differences between the effects on speech perception and memory for spoken words of talker and rate variability on the one hand, and amplitude variability on the other, is due to differences in their phonetic relevance to the listener.

A wider range of sources of variability needs to be investigated in order to provide conclusive evidence for one of these two alternative explanations for the different effects of different sources of variability. For example, it may be enlightening to investigate the effects of variations in dialect, vocal effort, emotional state and other such para-linguistic characteristics, as well as the effects of non-linguistic factors such as filtering characteristics due to different microphones or recording conditions. Nevertheless, so far as we can tell from the available data, all instance-specific stimulus attributes appear to be retained in memory to the extent that listeners are able to detect such changes. There is now a growing body of converging evidence demonstrating that the processes of speech perception and spoken word recognition operate in the context of highly detailed representations of the acoustic speech signal, rather than on idealized abstract symbolic representations of abstract linguistic information. We believe these are important new observations about speech and spoken language processing that have broad implications for future research and theory about speech perception.

References

American National Standards Institute (1971). *Method for measurement of monosyllabic word intelligibility*. (American National Standard S3.2-1960 [R1971]). New York: Author.

413

Church, B. A. & Schacter, D. L. (1994). Perceptual specificity of auditory priming: Implicit memory for voice intonation and fundamental frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* **20**, 521-533.

Craik, F. I. M. & Kirsner, K. (1974). The effect of speaker's voice on word recognition. *Quarterly Journal of Experimental Psychology,* **26**, 274-284.

Fant, G. (1973). *Speech sounds and features*. Cambridge, MA: MIT Press.

Goldinger, S. D., Pisoni, D. B., & Logan, J. S. (1991). On the nature of talker variability effects on recall of spoken word lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* **17**, 152-162.

Halle, M. (1985). Speculations about the representation of words in memory. In V. A. Fromkin (Ed.), *Phonetic linguistics* (pp. 101-114). New York, NY: Academic Press.

Johnson, K. (1990). The role of perceived speaker identity in F0 normalization of vowels. *Journal of the Acoustical Society of America,* **88**, 642-654.

Joos, M. A. (1948). Acoustic phonetics. *Language*, **24**, 1-136.

Kewley-Port, D. (1983). Time-varying features as correlates of place of articulation in stop consonants. *Journal of the Acoustical Society of America,* **73**, 322-335.

Klatt, D. H. (1973). Interaction between two factors that influence vowel duration. *Journal of the Acoustical Society of America,* **54**, 1102-1104.

Klatt, D. H. (1976). Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *Journal of the Acoustical Society of America,* **59**, 1208-1221.

Ladefoged, P. & Broadbent, D. E. (1957). Information conveyed by vowels. *Journal of the Acoustical Society of America,* **29**, 98-104.

Laver, J. (1989). Cognitive science and speech: A framework for research. In H. Schnelle and N. O. Bernsen (Eds.), *Logic and linguistics: Research directions for cognitive science. European Perspectives,* (pp. 37-70). Hillsdale, NJ: Erlbaum.

Laver, J. & Trudgill, P. (1979). Phonetic and linguistic markers in speech. In K. R. Scherer and H. Giles (Eds.), *Social markers in speech,* (pp. 1-32). Cambridge, UK: Cambridge University Press.

Lehiste, I. (1972). The timing of utterances and linguistic boundaries. *Journal of the Acoustical Society of America,* **51**, 2018-2024.

Luce, P. A. & Carrell, T. D. (1981). *Creating and editing waveforms using WAVES* (Research in Speech Perception, Progress Report No. 7). Bloomington, IN: Indiana University Speech Research Laboratory.

Martin, C. S., Mullennix, J. W., Pisoni, D. B., & Summers, W. V. (1989). Effects of talker variability on recall of spoken word lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **15**, 676-684.

Miller, J. L. (1987). Rate-dependent processing in speech perception. In A. Ellis (Ed.), *Progress in the psychology of language*, (pp. 119-157). Hillsdale, NJ: Erlbaum.

Miller, J. L., & Volaitis, L. E. (1989). Effects of speaking rate on the perceptual structure of a phonetic category. *Perception and Psychophysics*, **46**, 505-512.

Mullennix, J. W., Pisoni, D. B., & Martin, C. S. (1989). Some effects of talker variability on spoken word recognition. *Journal of the Acoustical Society of America*, **85**, 365-378.

Nearey, T. (1989). Static, dynamic, and relational properties in vowel perception. *Journal of the Acoustical Society of America*, **85**, 2088-2113.

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1995). Effects of stimulus variability on perception and representation of spoken words in memory. *Perception and Psychophysics*, **57**, 989-1001.

Palmeri, T. J., Goldinger, S. D., & Pisoni, D. B. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **19**, 309-328.

Peterson, G. E. & Barney, H. L. (1952). Control methods used in the study of vowels. *Journal of the Acoustical Society of America*, **24**, 175-184.

Picheny, M. A., Durlach, N. I. & Braida, L. D. (1986). Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech. *Journal of Speech and Hearing Research*, **29**, 434-446.

Picheny, M. A., Durlach, N. I. & Braida, L. D. (1989). Speaking clearly for the hard of hearing III: An attempt to determine the contribution of speaking rate to differences in intelligibility between clear and conversational speech. *Journal of Speech and Hearing Research*, **32**, 600-603.

Pisoni, D. B. (1993). Long-term memory in speech perception: Some new findings on talker variability, speaking rate and perceptual learning. *Speech Communication*, **13**, 109-125.

Pisoni, D. B. (1997). Some thoughts on "Normalization" in speech perception. In J. Mullennix and K. A. Johnson (Eds.), *Talker variability in speech processing*, (pp. 9-32). Academic Press.

Port, R. F. (1981). Linguistic timing factors in combination. *Journal of the Acoustical Society of America*, **69**, 262-274.

Schacter, D. L. & Church, B. A. (1992). Auditory priming: Implicit and explicit memory for words and voices. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **18**, 915-930.

Sheffert, S. M. & Fowler, C. A. (1995). The effects of voice and visible speaker change on memory for spoken words. *Journal of Learning and Memory*, **34**, 665-685.

Sommers, M. S., Nygaard, L. C., & Pisoni, D. B. (1994). Stimulus variability and spoken word recognition. I. Effects of variability in speaking rate and overall amplitude. *Journal of the Acoustical Society of America*, **96**, 1314-1324.

Stevens, K. N. & Blumstein, S. E. (1978). Invariant cues for place of articulation in stop consonants. *Journal of the Acoustical Society of America*, **64**, 1358-1368.

Summerfield, Q. (1981). On articulatory rate and perceptual constancy in phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance*, 7, 1074-1095.

Uchanski, R. M., Choi, S., Braida, L. M., Reed, C. M., & Durlach, N. I. (1996). Speaking clearly for the hard of hearing IV: Further studies of the role of speaking rate. *Journal of Speech and Hearing Research*, **39**, 494-509.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 21 (1996-1997)
Indiana University

Acoustic, Psychometric and Lexical Neighborhood Properties of the Spondaic Words: A Computational Analysis of Speech Discrimination Scores[1]

Ted A. Meyer,[2] David B. Pisoni, Paul A. Luce,[3] and Robert C. Bilger[4]

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405

---

[1] Research supported by NIH/NIDCD Grant DC00111 and Training Grant DC00012. The authors would like to thank Linnette A. Caldwell and Darla J. Sallee for clerical assistance with this project. The authors would like to thank Drs. Mario A. Svirsky, Karen I. Kirk, and Steven B. Chin for their comments on an earlier version of the manuscript.

[2] Indiana University School of Medicine, Department of Otolaryngology, Indianapolis, IN 46202

[3] State University of New York at Buffalo, Department of Psychology, Buffalo, NY 14260

[4] University of Illinois, Department of Speech & Hearing Science, Champaign, IL 61820

Acoustic, Psychometric and Lexical Neighborhood Properties of the Spondaic Words: A Computational Analysis of Speech Discrimination Scores

Abstract. Luce and Pisoni (in press) developed a model of spoken word recognition based on neighborhood probability characteristics of monosyllabic words. This model, known as the Neighborhood Activation Model (NAM), assumes that words are recognized in the context of other similar sounding patterns in the lexicon. NAM predicts that word recognition is dependent on how easy or difficult it is to confuse the stimulus word with phonetically similar words in a lexical neighborhood. Cluff and Luce (1990) demonstrated that spondaic word recognition is also dependent on the confusability of the individual syllables. Spondaic words with two easy to recognize syllables are identified more accurately than spondaic words with two hard to recognize syllables. Recently, Bilger et al. (submitted) analyzed the 36 spondaic words [Tillman et al. (1963) recording of the Hirsh et al. (1952) spondaic words] currently used to determine the speech reception threshold (SRT) (American Speech-Language-Hearing Association, 1988). They found that the thresholds and slopes of psychometric functions for the spondaic words were not equivalent, and that the spondees were not equally intelligible. In this study, we computed the lexical neighborhood characteristics (Pisoni et al., 1985) of the individual syllables in these spondaic words and compared these values to psychophysical and acoustic measures obtained from Bilger et al. Although spondaic word thresholds were not related to any of the lexical neighborhood measures of the words, the slopes of the psychometric functions were negatively correlated with the neighborhood densities of the spondaic words. This finding demonstrates that the rate at which a word becomes intelligible is inversely related to the confusability of the word with its lexical neighbors in memory. Implications for the development of new tests for speech discrimination are discussed.

Introduction

Recent work by Bilger (Bilger, Matthies, Meyer, & Griffiths, submitted; Meyer & Bilger, 1997) has examined the utility of spondaic words (words with two syllables having approximately equal stress) in speech discrimination tests. The spondaic words used to determine thresholds for speech were chosen specifically and modified for the original recording to be of equal intelligibility (Hirsh et al., 1952). In subsequent studies, the thresholds for the individual spondees recorded by Hirsh were not equivalent (Bowling & Elpern, 1961; Curry & Cox, 1966; Wilson & Margolis, 1983). Psychometric functions generated for the individual spondaic words recorded by Tillman (Tillman, Carhart & Wilber, 1963) also clearly demonstrate that the spondaic words are not equally intelligible over a range of presentation levels and that the slopes of psychometric functions for these words are not equivalent (Bilger et al., submitted; Young, Dudley & Gunter, 1982). These findings pose a serious problem for clinicians who routinely derive speech reception thresholds using the spondaic words as stimuli (Penrod, 1994). One possible solution to this problem is to use spondaic words with similar acoustic properties. However, Bilger et al. also found poor correlations between gross acoustic measures (durations and amplitudes) and psychophysical measures (thresholds and slopes of psychometric functions) of the individual spondaic words. Their findings suggest that other structural factors related to acoustic-phonetic discriminability may play a role in determining spondaic word recognition.

For the past ten years, Luce and his colleagues (Luce, 1986; Luce & Pisoni, in press; Luce, Pisoni, & Goldinger, 1990) have been developing a model of spoken word recognition known as the Neighborhood Activation Model (NAM). This model attempts to account for the effects of lexical activation and competition on speech perception and spoken word recognition. In particular, this model assumes that words are recognized relationally in the context of other phonetically similar words in the lexicon. A particular sound pattern activates multiple lexical items in memory and this pattern of activation affects lexical discrimination and subsequent recognition performance.

NAM predicts that when a spoken word is perceived, it activates a set of representations of similar sounding words in memory. This set of phonetically similar words is called a "lexical neighborhood." Once the neighborhood of the stimulus is activated, the word recognition system must then discriminate among the activated neighbors and decide which of the neighbors best matches the stimulus. The lexical discrimination and decision process is influenced by three factors: (1) the frequency of the target word; (2) the number of phonetically similar words (neighbors) activated in memory by the stimulus pattern; and (3) the frequencies of occurrence of the neighbors in the similarity neighborhood. Because of increased competition among activated items in memory, words from densely populated lexical neighborhoods are recognized more slowly and less accurately than words from sparsely populated neighborhoods. In addition, the presence of high frequency neighbors produces increased competition among words that are phonetically similar to the target word, thereby also reducing recognition performance.

According to NAM, identification performance can be predicted by the following rule:

$$p(ID) = \frac{p(Stim) * Freq_s}{p(Stim) * Freq_s + \sum_{j=1}^{n} p(Neighbor_j) * Freq_j}$$

where p(ID) is the probability of correction identification, $p(Stim)*Freq_s$ is the frequency-weighted probability of identifying the stimulus word based on acoustic-phonetic information, and $p(Neighbor_j)*Freq_j$ is the frequency-weighted probability of identifying a neighbor. According to this rule, NAM predicts that stimulus words with few low-frequency neighbors will be identified most accurately (because of the relatively small denominator in Equation 1, which indexes the degree of competition among neighbors activated in memory). NAM also predicts that stimulus words with many high-frequency neighbors will be identified least accurately (large denominator in Equation 1).

Luce and Pisoni (in press) performed numerous regression analyses to evaluate the effects of several variables in NAM on speech recognition scores for monosyllabic words at various signal-to-noise (S/N) ratios. Across all S/N ratios, Luce and Pisoni found that variables from the model explain between 3% and 23% of the variance in the subject response. They also found that as the signal was degraded (lower S/N ratio), the structure of the lexical neighborhood of the stimulus becomes more important (i.e., explained more variance). The relation between S/N ratio and the lexical neighborhood of the stimulus may be of particular importance in examining the signals used in speech discrimination tests in which the stimuli are presented at or near threshold.

In examining how NAM applies to more longer and more complex words, Cluff and Luce (1990) assessed the contribution of the individual syllables to spondaic word recognition. Cluff and Luce categorized the individual syllables of the spondaic words used in their experiment according to the lexical neighborhood properties of the syllables. An "easy" syllable was a high-frequency word in a low-frequency, low-density neighborhood, and a "hard" syllable was a low-frequency word in a high-frequency, high-density neighborhood. The frequencies of the words were obtained from the Kucera and Francis (1967) word counts. Cluff and Luce found that spondees with an "easy-easy" syllable pattern were the easiest to identify, whereas spondees with a "hard-hard" syllable pattern were the most difficult to identify. Identification of spondees with a single "easy" and a single "hard" syllable was of intermediate difficulty. Further, spondees with a "hard-easy" syllable pattern were easier to identify than spondees with an "easy-hard" syllable pattern. Cluff and Luce attributed this asymmetry in performance primarily to a "retroactive" pattern of influence in spoken word recognition, a "hard" second syllable influenced perception more than an "easy" second syllable. In other words, compound words in English are not recognized strictly left to right in serial order.

In addition to providing a better fundamental understanding of how listeners with normal hearing are able to recognize spoken words, the fundamental assumptions of NAM also have several important clinical implications in terms of measuring speech perception in listeners with hearing impairments. For example, Kirk and her colleagues (Kirk, Pisoni, & Osberger, 1995; Kirk, Pisoni, Sommers, Young, & Evanson, 1995) have recently developed a new class of speech recognition tests based on assumptions of NAM for subjects with impaired hearing and cochlear implants (CIs). Their findings show that hearing-impaired listeners and patients with CIs also recognize words in the context of other phonetically similar words in their lexicons. And, moreover, they recognize words by reference to "similarity neighborhoods" through a two-step process of bottom-up "activation" followed by "lexical discrimination," as assumed by NAM (see Luce & Pisoni, in press). Differences in speech perception performance between listeners with normal hearing and impaired hearing can be accounted for by NAM which makes specific predictions about the effect of frequency and acoustic-phonetic similarity and how these factors influence recognition performance.

In this paper, we report the results of several analyses that compared the acoustic and psychophysical measurements on the Tillman et al. (1963) recording of spondaic words (Bilger et al., submitted; Meyer & Bilger, 1997) to the lexical properties of the individual syllables of the spondaic words. We also examine the relation between threshold measures of intelligibility (Bowling & Elpern, 1961; Curry & Cox, 1966; Wilson & Margolis, 1983) of the Hirsh et al. (1952) recording of the spondaic words and the lexical neighborhood properties of the individual syllables of the words. Using computational analyses of these sound patterns, we hoped to gain a better understanding of the structural factors that influence spondaic word recognition near threshold.

Methods

Thresholds of Intelligibility

Hirsh et al. (1952) recorded 36 spondaic words to be used as stimuli for obtaining speech reception thresholds (SRTs). The words were intended to be of equal stress on the two syllables, and of equal intelligibility. Bowling and Elpern (1961) and Curry and Cox (1966) measured the average sensation level at which the individual words first became intelligible to a large group of normal-hearing listeners. Both studies determined that the spondaic words were not of equal intelligibility and suggested that smaller "more homogeneous" sets of spondaic words should be used for determining SRTs. Bowling and Elpern (1961) suggested a shorter list of 23 spondees, whereas Curry and Cox suggested that 27 of their words were "equally intelligible." In these two studies, the words in the "equally intelligible" lists had thresholds that were within one standard deviation of the mean threshold of the entire list of spondees. Wilson and

4-6

Margolis (1983) constructed a separate list of 21 “equally intelligible” spondaic words by averaging the results from Bowling and Elpern and Curry and Cox and adding a correction factor to account for the different presentation levels of the different words. In the Wilson and Margolis study, the words in the “equally intelligible” list also had thresholds that were within one standard deviation of the mean threshold of the entire list of spondees.

In the present study, we examined the relation between the individual word thresholds from the entire spondaic word lists from the three studies mentioned above and the lexical neighborhood properties of the words. As was stated earlier, NAM predicts that the lexical neighborhood of a word is of greater importance to word recognition as the S/N ratio becomes more degraded. Our goal, in this study, was to determine whether any of the lexical neighborhood properties of the words might play a role in predicting the performance of subjects in identifying a word when it first becomes intelligible. In other words, do “easy” words (high-frequency words from sparse neighborhoods) have lower thresholds than “hard” words (low-frequency words from dense neighborhoods)?

$$P(C)=\frac{1}{e^{(-mX_i-b)}}$$

Psychometric Functions

Young et al. (1982) generated psychometric functions for the individual spondaic words (Tillman et al., 1963 recording). They fit lines to the middle (linear) portion of the psychometric functions to estimate the thresholds (50% points) and slopes of the functions. Bilger et al. (submitted) also generated psychometric functions for the spondaic words from the data of Young et al. (1982), but in a slightly different manner. In Bilger et al., the data were fitted to the logit function described as follows:

in which *P(C)* is the probability of a correct response, and *m* is the slope of the ogive at its midpoint, *b* [P(C) = .50] (Lord, 1980). This function was fit to a least squares criterion using the logit transformation of proportion of correct responses as one variable in the calculation of a linear regression with intensity level. In the present study, the thresholds [P(C) = .50] and slopes of the psychometric functions of the individual spondaic words estimated by Bilger et al. (submitted) were compared to acoustic and lexical neighborhood measures of the individual syllables of the words.

Acoustic Measures

Bilger et al. (submitted) also measured acoustic parameters of the spondaic words. A tape-recorded version of the Tillman et al. (1963) recording of the spondaic words from CID W-1 (Hirsh et al., 1952) was digitized with 16-bit intensity resolution at a rate of 10-kHz (Data Translation, Model 2823). To eliminate artifacts, the stimuli were played through a bandpass filter (100-4000 Hz) (Wavetek-Rockland, Model 751-A) prior to the A/D conversion. The digitized spondees were then analyzed using Interactive Laboratory System (ILS-PC) software. The measures obtained were the RMS amplitude of each syllable and the overall RMS amplitude of each spondee (in dBV), as well as the duration of each syllable and the total duration of each spondee (in ms). These values were compared to the psychometric and lexical neighborhood measures of the individual syllables of the spondaic words.

BEST COPY AVAILABLE

Lexical Neighborhood Measures

Computational analyses of the individual spondaic words were carried out using a computerized database. The lexical database consists of 20,000 words from the *Webster's Pocket Dictionary*. Details of the development of the database are described elsewhere (Luce, 1986; Luce & Pisoni, in press; Nusbaum, Pisoni, & Davis, 1984; Pisoni, Nusbaum, Luce, & Slowiaczek, 1985). The word frequencies were obtained from Kucera and Francis (1967). Lexical neighborhoods for a given word were computed by using a single-phoneme substitution, addition, or deletion rule (Greenberg & Jenkins, 1964; Landauer & Streeter, 1973). For example, if the target word is "art," the lexical neighborhood of "art" would include words such as "ark" (substitution), "dart" (addition), and "are" (deletion). In the present study, we obtained the frequency of occurrence for the individual syllables of the spondaic words as well as the frequency of occurrence and density of the lexical neighborhoods of the syllables. The obtained values for the two syllables were then averaged to obtain a measure for the entire spondaic word. These values were compared to the psychometric and acoustic properties of the spondees.

Statistical Analysis

The SigmaStat statistical software package was used to perform a multiple correlational analysis between the psychometric, acoustical, and lexical neighborhood properties of the spondaic words. The results are reported below.

Results

Thresholds of Intelligibility

The thresholds obtained from the Hirsh et al. (1952) recording of the spondaic words (Bowling & Elpern, 1961; Curry & Cox, 1966; Wilson & Margolis, 1983) are shown in Table I. For the three different studies, the range of the levels at which the individual spondaic words became intelligible was quite large (10.0 dB - BE, 8.1 dB - CC, and 7.6 dB - WM) suggesting that the individual spondaic words are not equally intelligible. The correlations between the spondaic word thresholds obtained from the three studies are quite high (BE x CC, $r = .769$, $p < .001$; BE x WM, $r = .763$, $p < .001$; CC x WM, $r = .611$, $p < .001$) (see Table IV) and suggest that the word thresholds are strongly related. Although the correlations between the levels at which the spondaic words first become intelligible in the three studies are significantly greater than zero, the three "homogeneous lists" generated by the three studies described earlier cannot be considered "equivalent" lists of spondees (for complete details see Bilger et al., submitted). The three short lists appear quite similar, but appearances can be deceptive. The appropriate statistic to describe the similarity between two word lists (nominal variables) is the contingency coefficient r (Stevens, 1951; Welkowitz, Ewen, & Cohen, 1991). Only 20 of the words in the 23-word list generated by Bowling and Elpern appear in the 27-word list of Curry and Cox ($r = .367$, $p = .0275$), only 17 of Bowling and Elpern's 23 "homogeneous" words appear in the 21-word list of Wilson and Margolis ($r = .420$, $p = .012$), while only 17 of the words in the Curry and Cox list appear in the Wilson and Margolis list ($r = .293$, $p = .0790$). Although the contingency correlations between the BE and CC and the BE and WM lists are significantly greater than 0.0 ($p < .05$), the values of r are also well below any reasonable criterion for equivalence. In addition, only 14 of the 36 spondees appear on all three "homogeneous lists" (see Table I). Thus, the individual word thresholds are more closely related than the "homogeneous" lists created from these thresholds. The relation between the individual word thresholds and the lexical neighborhood values will be discussed later.

Table I.

Spondaic word thresholds of intelligibility (in dB re the lowest level at which a word was intelligible) from Bowling and Elpern (1961), Curry and Cox (1966), and Wilson and Margolis (1983) (Hirsh as talker). Superscripts after the words indicate inclusion in a "homogeneous" list, 1-BE, 2-CC, 3-WM.

Word	BE	CC	WM
AIRPLANE^{1,2}	4.7	4.2	3.06
ARMCHAIR^{1,2,3}	7.5	4.4	8.25
BASEBALL	3.0	3.0	4.18
BIRTHDAY^{1,2}	6.3	5.6	3.87
COWBOY^{1,2,3}	4.3	5.0	5.13
DAYBREAK^{2}	8.3	6.6	9.20
DOORMAT^{2}	8.3	6.8	10.61
DRAWBRIDGE^{2,3}	8.2	4.8	6.43
DUCKPOND^{3}	8.3	10.1	4.78
EARDRUM1^{2,3}	6.9	6.1	6.70
FAREWELL^{1,2,3}	6.3	8.0	7.03
GRANDSON	12.2	8.4	10.62
GREYHOUND^{1,2,3}	7.0	6.0	7.12
HARDWARE^{2,3}	3.5	4.9	4.54
HEADLIGHT	8.3	8.6	9.52
HORSESHOE^{1,3}	7.8	8.6	7.07
HOTDOG^{1}	5.0	3.8	2.98
HOTHOUSE	12.7	10.9	9.87
ICEBERG^{1,2}	4.8	4.0	3.01
INKWELL^{1,2,3}	8.8	7.0	7.52
MOUSETRAP^{1,2,3}	7.7	7.7	7.95
MUSHROOM^{2}	10.3	7.7	9.81
NORTHWEST^{1,2}	5.2	6.2	3.86
OATMEAL^{1,2,3}	7.0	7.6	5.82
PADLOCK^{2}	8.3	8.0	8.93
PANCAKE^{1,3}	7.7	8.0	7.32
PLAYGROUND^{1,2}	4.5	4.6	3.91
RAILROAD^{1,2,3}	4.8	6.5	6.02
SCHOOLBOY^{2,3}	8.5	6.9	6.95
SIDEWALK^{1,2,3}	6.0	6.5	5.31
STAIRWAY^{1,2,3}	6.3	8.0	8.71
SUNSET^{1,2,3}	6.2	6.2	5.32
TOOTHBRUSH^{1,2,3}	7.5	8.0	7.35
WHITEWASH^{1,2,3}	7.2	7.8	5.19
WOODWORK^{1,2,3}	4.3	4.8	7.66
WORKSHOP	2.7	2.8	3.82
Mean	6.84	6.50	6.54
S. D.	2.27	1.91	2.25

Psychometric Functions

Midpoints (thresholds in dB SPL) and slopes (in %/dB) of the psychometric functions from the Bilger et al. (submitted) analysis of the spondaic words from Young et al. (1982) are shown in Table II. Thresholds ranged from 15.0 to 21.6 dB SPL with a mean of 18.9 dB SPL. The slopes of the 36 psychometric functions ranged from 6.8 to 16.9 %/dB with a mean of 10.7 %/dB. Although the spondees are supposed to be "equally intelligible," the average thresholds span a range of nearly 7 dB, and the rate at which the words become intelligible varies over a range of 10%/dB. Below we examine whether the acoustic measures of the words relate to these psychometric measures and determine whether NAM plays a role in the identification of the spondees at threshold.

Acoustic Measures

Acoustic measures of overall amplitude and duration of the spondaic words are also shown in Table II. The duration of the spondaic words ranged from 723 to 992 ms with a mean of 847 ms. The mean duration of the first syllable was shorter (315 ms) than the mean duration of the second syllable (486 ms), and 18 of the 36 spondees have silent periods (closures) between the syllables in the Tillman recording. The RMS levels of the spondaic words ranged from -13.5 to -11.3 dBV with a mean of -12.3 dBV. Although shorter in duration, the first syllables had a greater mean amplitude (-11.2 dBV) than the second syllables (-12.7 dBV). The small ranges and standard deviations of the amplitudes of the individual syllables as well as the entire words suggest that by monitoring their utterances with a VU meter, Tillman et al. were able to produce a list of spondees that were quite similar in overall amplitude. We will examine how these acoustic measures relate to psychometric and the lexical neighborhood characteristics of the spondaic words.

Lexical Neighborhood Measures

Lexical neighborhood measures for the 36 spondaic words based on computational analyses of the similarity spaces from the lexical database are shown in Table III. Included in Table III are the frequency of occurrence for the individual syllables as well as the frequency of occurrence of the entire word. The frequency of occurrence of the words in the individual syllable's lexical neighborhood as well as the density of the individual syllable's lexical neighborhood are also included. As described above, values for the spondaic words were generated by computing the arithmetic average of the measures for the individual syllables. Several of the individual syllables that make up the spondees are "easy" words, others are "hard" words, and still others are of intermediate difficulty. That is, some words come from lexical neighborhoods where there is little competition from other phonetically similar words (i.e., "easy" words), whereas other words come from lexical neighborhoods where there is much more competition from other phonetically similar words (i.e., "hard" words). As shown in Table III, there is a great deal of variability in the lexical neighborhood measures for the different spondees.

Statistical Analysis

A multiple correlational analysis was performed to assess the relations between the different lexical neighborhood, acoustic, and psychophysical measures of the spondaic words described above. The correlations are shown in Table IV. Several interesting findings emerged from the correlational analysis. First, the spondaic word thresholds (levels at which the words first became intelligible) obtained from the studies with Hirsh as talker (Bowling & Elpern, 1961; Curry & Cox, 1966; Wilson & Margolis, 1983) were not significantly correlated with any of the lexical neighborhood measures for the individual syllables

424

Table II.

Midpoints [dB SPL for a P(C)=50%] and slopes (%/dB) from two-parameter logistic model fitted to the data of Young et al. (1982) (Tillman as talker), and word duration and RMS levels (from Bilger et al., submitted). 1st Syl - first syllable, 2nd Syl - second syllable, dBV = dB re 1V.

WORD	Midpoint [P(C) = 50%] (dB SPL)	Slope (%/dB) Total	Duration 1st Syl (ms)	Duration 2nd Syl (ms)	Duration Total	RMS level 1st Syl (dBV)	RMS level 2nd Syl (dBV)	RMS level (dBV)
AIRPLANE	18.1	10.8	858	262	506	-12.4	-10.7	-12.8
ARMCHAIR	17.7	9.8	858	378	461	-11.8	-11.8	-11.8
BASEBALL	17.4	11.5	858	307	416	-12.6	-12.1	-11.6
BIRTHDAY	18.8	9.0	813	243	416	-11.3	- 9.9	-10.8
COWBOY	18.0	11.8	723	307	416	-11.5	-11.2	-11.8
DAYBREAK	20.7	8.6	666	378	288	-11.6	-11.7	-11.4
DOORMAT	19.7	13.2	640	352	288	-11.5	-10.9	-12.4
DRAWBRIDGE	18.5	12.1	884	378	506	-12.5	-11.5	-13.4
DUCKPOND	18.9	10.4	896	154	550	-12.7	- 9.0	-12.8
EARDRUM	17.7	7.0	884	378	506	-12.4	-11.9	-12.9
FAREWELL	20.8	7.7	922	397	525	-11.7	-11.1	-12.2
GRANDSON	19.2	9.2	940	442	416	-11.6	-10.1	-11.9
GREYHOUND	21.7	9.1	992	397	595	-12.9	-12.4	-13.3
HARDWARE	17.3	11.6	819	333	486	-12.7	-12.7	-12.7
HEADLIGHT	18.6	7.8	768	243	525	-12.7	-10.5	-14.3
HORSESHOE	17.5	13.0	813	307	506	-12.2	-11.3	-12.9
HOTDOG	15.0	10.9	896	198	506	-12.8	- 9.9	-12.9
HOTHOUSE	18.4	10.1	858	243	570	-13.3	-10.7	-14.6
ICEBERG	16.8	15.0	940	333	525	-12.7	-12.2	-12.4
INKWELL	18.5	6.8	877	243	570	-11.8	-10.3	-12.0
MOUSETRAP	20.3	9.5	896	378	461	-12.5	-11.6	-12.7
MUSHROOM	20.0	7.8	877	333	506	-12.6	-11.9	-12.8
NORTHWEST	18.8	10.8	941	307	550	-13.1	-11.4	-13.6
OATMEAL	21.6	10.4	813	198	486	-11.7	- 8.8	-12.4
PADLOCK	19.7	10.3	793	307	486	-12.3	-11.4	-13.2
PANCAKE	20.1	9.3	858	378	461	-12.3	-11.5	-12.9
PLAYGROUND	17.8	9.9	973	378	595	-12.6	-11.9	-13.1
RAILROAD	19.2	8.9	948	442	506	-11.7	-11.6	-11.9
SCHOOLBOY	20.3	16.9	896	480	416	-12.4	-12.7	-12.1
SIDEWALK	19.4	11.8	793	307	486	-12.3	-11.4	-13.0
STAIRWAY	18.8	10.0	839	378	461	-11.8	-11.4	-12.2
SUNSET	21.3	12.4	813	352	461	-12.1	-11.2	-13.0
TOOTHBRUSH	18.9	14.9	723	134	486	-13.5	-12.4	-13.0
WHITEWASH	16.5	11.6	857	224	525	-12.8	-10.1	-13.4
WOODWORK	19.2	11.9	768	262	506	-12.9	-11.3	-14.1
WORKSHOP	17.9	12.8	813	224	525	-12.7	-10.2	-13.8
Mean	18.86	10.68	847	315	486	-12.3	-11.2	-12.7
S. D.	1.47	2.25	80	82	68	.55	.94	.82

Table III.

Lexical neighborhood properties of the spondaic words. Word values were obtained by averaging the word frequencies and neighborhood properties of the individual syllables. Syl1 -1st syllable, Syl2 -2nd syllable, Freq -frequency, Nhbd -neighborhood, Den -density.

Word	Mean Word Freq	Syl1 Freq	Syl2 Freq	Nhbd Freq	Syl1 Nhbd Freq	Syl2 Nhbd Freq	Nhbd Den	Syl1 Nhbd Den	Syl2 Nhbd Den
AIRPLANE	214	266	162	337.02	560.83	113.20	20.0	30.0	10.0
ARMCHAIR	80	94	66	443.09	559.22	326.95	14.5	9.0	20.0
BASEBALL	101	91	110	105.69	41.63	169.75	25.5	27.0	24.0
BIRTHDAY	378	70	686	197.63	67.56	327.70	21.5	16.0	27.0
COWBOY	135	29	242	523.53	121.37	925.69	16.0	19.0	13.0
DAYBREAK	388	686	90	171.74	327.70	15.77	20.0	27.0	13.0
DOORMAT	160	312	8	325.02	14.00	636.03	21.5	13.0	30.0
DRAWBRIDGE	77	56	98	37.75	42.00	33.50	6.0	6.0	6.0
DUCKPOND	17	9	25	27.58	41.88	13.29	16.0	25.0	7.0
EARDRUM	20	29	11	1111.19	1774.39	448.00	20.5	31.0	10.0
FAREWELL	491	84	897	359.81	614.43	105.19	29.5	28.0	31.0
GRANDSON	163	48	278	172.00	37.57	306.42	16.5	7.0	26.0
GREYHOUND	43	80	7	104.72	57.56	151.89	13.5	18.0	9.0
HARDWARE	121	202	39	191.60	22.44	360.76	21.5	18.0	25.0
HEADLIGHT	379	424	333	255.04	354.80	155.29	30.0	25.0	35.0
HORSESHOE	68	122	14	889.80	76.27	1703.33	17.5	11.0	24.0
HOTDOG	103	130	75	147.03	282.18	11.88	18.0	28.0	8.0
HOTHOUSE	361	130	591	202.16	282.18	122.14	17.5	28.0	7.0
ICEBERG	23	45	1	219.00	395.31	42.69	14.5	16.0	13.0
INKWELL	452	7	897	72.24	39.29	105.19	22.5	14.0	31.0
MOUSETRAP	15	10	20	46.64	79.43	13.85	13.5	14.0	13.0
MUSHROOM	193	1	384	43.22	67.40	19.04	19.0	15.0	23.0
NORTHWEST	221	206	235	81.88	87.75	76.00	11.5	4.0	19.0
OATMEAL	15	1	30	479.79	882.36	77.21	26.5	25.0	28.0
PADLOCK	15	8	23	150.36	225.12	75.61	28.5	26.0	31.0
PANCAKE	15	16	13	202.81	301.35	104.27	28.5	31.0	26.0
PLAYGROUND	193	200	186	107.11	75.63	138.60	10.5	16.0	5.0
RAILROAD	127	16	237	34.32	31.39	37.24	32.5	36.0	29.0
SCHOOLBOY	367	492	242	474.22	22.75	925.69	10.5	8.0	13.0
SIDEWALK	240	380	100	113.21	122.35	104.07	19.0	23.0	15.0
STAIRWAY	465	16	913	232.21	27.75	436.67	21.0	12.0	30.0
SUNSET	346	278	414	228.85	306.42	151.28	29.0	26.0	32.0
TOOTHBRUSH	32	20	44	1031.83	2057.86	5.80	9.5	14.0	5.0
WHITEWASH	201	365	37	183.64	302.27	65.00	9.0	11.0	7.0
WOODWORK	1765	2769	760	149.93	252.87	47.00	17.5	15.0	20.0
WORKSHOP	411	760	63	43.69	47.00	40.38	18.0	20.0	16.0
MEAN	233	235	231	263.82	294.51	233.12	19.1	19.2	18.9
S. D.	303	475	281	265.33	449.27	342.83	6.6	8.3	9.5

Table IV.

Correlation matrix between acoustical, psychophysical and lexical neighborhood variables (~ p < .05, * p < .01). wd - word, s1 - 1st syllable, s2 - 2nd syllable, nd - neighborhood, dur - duration, freq - frequency, den - density, rms - root mean square level, BE - Bowling & Elpern, CC - Curry & Cox, WM - Wilson & Margolis, Bilg - Bilger et al.

	BE	CC	WM	mean Bilg	slope Bilg	wd dur	s1 dur	s2 dur	wd rms	s1 rms	s2 rms	wd freq	s1 freq	s2 freq	nd freq	s1 nd freq	s2 nd freq	nd den	s1 nd den	s2 nd den
BE																				
CC	.769*																			
WM	.763*	.611*																		
mean Bilg		.387~	.440*																	
slope Bilg																				
wd dur																				
s1 dur																				
s2 dur						.649*														
wd rms							.388~													
s1 rms							-.566*													
s2 rms						-.576*		-.578*	.755*											
wd freq																				
s1 freq										-.338~										
s2 freq					-.335~															
wd nd freq																				
s1 nd freq																				
s2 nd freq																				
wd nd den					-.427*				.433*											
s1 nd den					-.348~															
s2 nd den									.560*					.406~						

or with those generated for the words. Second, the thresholds (levels at which the words were intelligible 50% of the time) of the psychometric functions generated by Bilger et al. (submitted) from Young et al.'s (1982) data with Tillman as talker were also not significantly correlated with any of the lexical neighborhood measures. Although NAM predicts that easy words (high-frequency words from sparse lexical neighborhoods) should be easier to recognize than hard words (low-frequency words from dense lexical neighborhoods), this relation does not appear to hold for the spondaic words presented at threshold (defined as either level of initial recognition or the 50% point on a psychometric curve). Third, none of the acoustic measures (RMS level, duration) of the individual syllables or the entire spondaic words were significantly related to the thresholds or slopes of the psychometric functions of the individual words. However, two significant correlations between the acoustic measures of the spondaic words and their lexical neighborhood characteristics emerged from this analysis. First, the overall RMS level of the word was positively correlated to the neighborhood density of the word ($r = +.433$, $p < .01$), and second, the RMS level of the first syllable was negatively correlated to the frequency of the first syllable ($r = -.388$, $p < .05$). These may represent random occurrences, because it is difficult to imagine any systematic relation between these measures especially given that the RMS measures were obtained from single utterances of the spondaic words.

Finally, the slopes of the psychometric functions generated by Bilger et al. were significantly correlated with the lexical neighborhood densities of the first syllable of the spondees ($r = -.348$, $p < .05$) as well as the lexical neighborhood densities generated for the whole spondaic words ($r = -.427$, $p < .01$). These negative correlations suggest that the rate at which a spondaic word (target) becomes recognized is inversely related to the number of words that are phonetically similar to the target. That is, words from dense neighborhoods are recognized at a slower rate than words from sparse neighborhoods. This finding would be anticipated based on the assumptions of NAM. From Equation 1, the probability of correctly identifying a stimulus is inversely related to the density (i.e., similarity) of the lexical neighborhood of the stimulus. If a stimulus lies in a dense lexical neighborhood, the denominator of Equation 1 is relatively large compared to the denominator when the stimulus lies in a sparse lexical neighborhood. Thus, the predicted change in the ability of a subject to correctly identify a stimulus (slope of a psychometric function) should be less when the stimulus is from a dense neighborhood (large denominator) than when the stimulus is from a sparse neighborhood (small denominator). This is precisely the result we observed here.

Discussion

Spoken language is infinitely variable. Each repetition of a word by an individual talker is a unique representation of speech, yet even under listening conditions that are far from ideal, we are able to recognize familiar words given that they are represented in our mental lexicons. Although the precise neural representation of how sound patterns of words are stored in one's mental lexicon is not known, and is currently a topic of great interest, the lexicon plays an important role in speech perception and spoken word recognition (Luce et al., 1990; Miller, 1946; Treisman, 1978a,b). In identifying an incoming stimulus, we compare that stimulus to a multidimensional representation of words stored in a psychological space. Several recent models of spoken word recognition have attempted to explain the process of word recognition. In general, the models compare information from the incoming stimulus to information stored in a subject's mental lexicon (see Lively, Pisoni, & Goldinger, 1994, for a review).

NAM is a model of spoken word recognition containing both bottom-up and top-down processing components. As the incoming stimulus is processed, a set of phonetically similar words, a "lexical neighborhood," or set of word hypotheses is activated in memory. The density and frequency of the words in a lexical neighborhood have been shown to produce strong effects on the identification of the stimulus word (Luce et al., 1990). Furthermore, the lexicon may be activated and comparisons made to the incoming stimulus multiple times depending on the complexity of the stimulus. For example, in a series of experiments in which the stimuli were spondaic words, Cluff and Luce (1990) and Charles-Luce, Luce, and Cluff (1990) demonstrated that word recognition was dependent upon the principles of multiple activation and delayed commitment. They found that the lexical difficulty of the second syllable influenced the identification of the spondaic words (retroactive influence of the second syllable). That is, words with easy second syllables were identified a greater percentage of the time than words with hard second syllables. Also, words with a "hard-easy" syllable pattern were more easily recognized than words with an "easy-hard" syllable pattern. Their findings also showed that the "easy" second syllable allowed the "hard" first syllable to be recognized more readily, whereas the "hard" second syllable interfered with the recognition of the "easy" first syllable.

In addition to the contribution of the lexicon on speech perception, the specific information processing task used to measure speech discrimination has been shown to have a substantial impact on a listener's performance. For example, open-set and closed-set tasks place very different demands on the listener. Closed-set tests are essentially tests of speech pattern discrimination, whereas open-set tests involve activation, search and retrieval of a lexical representation from memory before a response can be made. Recently, Sommers, Kirk, and Pisoni (1997) found that open-set word recognition was more difficult for all their subjects when the stimuli were produced with multiple talkers instead of a single talker. Open-set recognition scores were also lower when the words were lexically "hard" than when the words were lexically "easy". In contrast, closed-set word recognition did not show effects of talker variability or lexical difficulty when words in the set were chosen to be easily confused with the target words. Sommers et al. concluded that closed-set tests may not adequately simulate the cognitive demands that individuals face in everyday listening situations. Furthermore, speech discrimination scores obtained using closed-set tests are not equivalent to scores obtained using open-set tests and therefore may not be able to predict speech perception performance in real-world conversational situations. In closed-set tests, where listeners simply have to discriminate differences among sound patterns, rather than recognize or identify words, the set of potential responses is no longer determined by the structure of the mental lexicon (Black, 1957). Listeners can restrict their lexical search and decision strategies to the available response alternatives provided in the forced-choice set instead of having to search through phonetically similar words in memory. In contrast, in open-set speech discrimination tests, listeners must compare the incoming speech signal to patterns stored in long-term lexical memory. Thus, the organization of words within the mental lexicon can have a significant influence on speech perception and spoken word recognition performance.

Summary and Conclusions

In the present study, we used a correlational analysis to examine relations between the acoustic and psychophysical measures of spondaic words and the lexical neighborhood properties of the individual syllables of these words. The results revealed several interesting findings. First, we found that the slopes of the psychometric functions generated by Bilger et al. (submitted) from the data of Young et al. (1982) were inversely related to the neighborhood densities of the spondaic words. Words from sparse lexical neighborhoods where there is little competition from other phonetically similar words become intelligible at a faster rate than words from dense neighborhoods where there is much greater competition. This pattern of results supports a major prediction of NAM -- that spoken words are recognized in relation to other words in the lexicon that have similar acoustic-phonetic properties.

Second, neither word frequency, neighborhood frequency, neighborhood density, nor any acoustical measures were significantly correlated with the spondaic word thresholds. This finding does not directly support the assumptions of NAM, however, the failure of the lexical neighborhood measures to correlate with spondaic thresholds may indicate that there is simply too little useful acoustic-phonetic information in the signal at threshold to strongly engage lexical representations in the recognition process. Performance at threshold may be more indicative of subjects' off-line guessing strategies based on very impoverished input than of activated lexical representations in memory. Another possibility is that identification of stimuli at threshold levels is almost entirely driven by the minimal amount of acoustic-phonetic information available to the listener. Because the present method of calculating similarity neighborhoods fails to take into account all but categorical changes in phonemes, this metric may fail to capture subtle phonemic confusions that are primarily responsible for spondee identification at threshold (see Luce and Pisoni, in press).

Another explanation for the lack of correlation for the thresholds may simply be that there is little variance to account for in the first place. Indeed, the standard deviation of the thresholds was only 1.47 dB (Bilger et al. psychometric analysis in Table II), which is much smaller than the standard deviation for the slopes (2.25 %/dB), which did produce significant correlations with the lexical variables. Whatever the precise explanation, Luce and Pisoni (in press) obtained similar results for words embedded in noise over a range of S/N ratios: Correlations between the lexical variables and identification accuracy were quite low at the lowest S/N ratios, and lexical effects were only found at signal levels above threshold.

The data from the present study and other recent investigations of the role of the lexicon in speech perception suggest a need for new theoretically-based tests to assess SRTs and other measures of speech recognition in clinical populations. Recent findings suggest that different populations such as the hearing impaired or patients with CIs may not identify spoken words in the same manner as listeners with normal hearing (see Koch, Carrell, Tremblay, & Kraus, 1996). The peripheral encoding of the speech signal is certainly very different for cochlear implant users than for normal-hearing subjects. And, it is very likely that these patients have different lexical representations of words as well. Different groups of subjects may use different equivalence classes than normal-hearing listeners do depending on the nature of their hearing impairments. These differences in their perceived similarity spaces may prove to be an important set of dimensions in assessing how cochlear implants and hearing aids work, and in measuring changes in speech discrimination performance over time using different aural rehabilitation and perceptual training methods.

New speech discrimination tests will need to embrace theoretical concepts and assumptions that specify and describe how normal-hearing listeners recognize spoken words and how they access information about the sound patterns of words from representations in long-term lexical memory (see Luce & Pisoni, in press). The process of "lexical discrimination," that is, how listeners recognize and identify spoken words is not simply equivalent to phoneme or nonsense syllable perception (Bilger & Wang, 1976; Miller & Nicely, 1955; Tyler & Moore, 1992; Wang & Bilger, 1973) where the primary emphasis is on speech feature recognition (see Rabinowitz, Eddington, Delhorne, & Cuneo, 1992). In lexical discrimination tests, listeners are asked to identify a sound pattern as a unique entry from among the words of their language. Successful completion of this task requires not only access to the sensory-based acoustic-phonetic information in the speech waveform but also retrieval of other structural information about the relationship of the sound pattern of the target word to phonetically similar words in the listener's lexicon. According to this approach, spoken words are recognized in the context of other similar words that the listener knows in his/her language. The present findings demonstrate that the large degree of variability typically observed among different words used in speech discrimination tests derives, in part, from the "lexical" and structural properties of the sound patterns of words, not only their acoustic features or correlates. These results therefore provide additional support for several of the major theoretical assumptions of the Neighborhood Activation Model of spoken word recognition. Furthermore, these results suggest a clinically useful line of research for future studies of hearing-impaired listeners where the stimulus materials are constructed in a theoretically motivated way based on what is currently known about the structural acoustic-phonetic properties of spoken words and how they are recognized by normal-hearing listeners. The findings from the present analysis demonstrate that listeners use information about the structural patterns of words in their lexicons to perceive speech over a wide range of S/N ratios. Moreover, the rate at which information about words is encoded or perceived in open-set speech intelligibility or discrimination tests is inversely related to the acoustic-phonetic confusability of the patterns in lexical memory. This important new finding is predicted by NAM which assumes that spoken words are recognized *relationally* in the context of other phonetically similar words in the language.

References

American Speech-Language-Hearing Association. (1988). Guidelines for determining threshold level for speech. *Asha, 30*, 85-88.

Bilger, R. C., Matthies, M. L., Meyer, T. A., & Griffiths, S. K. (submitted). Psychometric equivalence of recorded spondaic words as test items. *Journal of Speech-Language-Hearing Research*.

Bilger, R. C., & Wang, M. D. (1976). Consonant confusions in patients with sensorineural hearing loss. *Journal of Speech and Hearing Research, 19*, 718-748.

Black, J. W. (1957). Multiple choice intelligibility tests. *Journal of Speech and Hearing Disorders, 22*, 213-235.

Bowling, L. S., & Elpern, B. S. (1961). Relative intelligibility of items on CID Auditory Test W-1. *Journal of Auditory Research, 1*, 152-157.

Charles-Luce, J., Luce, P. A., & Cluff, M. S. (1990). Retroactive influence of syllable neighborhoods. In G. T. M. Altmann, (Ed.), *Cognitive Models of Speech Processing: Psycholinguistic and Computational Perspectives* (pp. 173-184). Cambridge: MIT Press.

Cluff, M. S. & Luce, P. A. (1990). Similarity neighborhoods of spoken two-syllable words: Retroactive effects on multiple activation. *Journal of Experimental Psychology: Human Perception and Performance, 3*, 551-563.

Curry, E. T., & Cox, B. P. (1966). The relative intelligibility of spondees. *Journal of Auditory Research, 6*, 419-424.

Greenberg, J. H., & Jenkins, J. J. (1964). Studies in the psychological correlates of the sound system of American English. *Word, 20*, 157-177.

Hirsh, I. J., Davis, H., Silverman, S. R., Reynolds, G., Eldert, E., & Benson, R. W. (1952). Development of materials for speech audiometry. *Journal of Speech and Hearing Disorders, 17*, 321-337.

Kirk, K. I., Pisoni, D. B., & Osberger, M. J. (1995). Lexical effects on spoken word recognition by pediatric cochlear implant users. *Ear & Hearing, 16*, 470-481.

Kirk, K. I., Pisoni, D. B., Sommers, M. S., Young, M., & Evanson, C. (1995). New directions for assessing speech perception in persons with sensory aids. *Annals of Otology, Rhinology, and Laryngology, 104* (Suppl. 166), 300-303.

Koch, D. B., Carrell, T. D., Tremblay, K., & Kraus, N. (1996). Perception of synthetic syllables by cochlear-implant users: Relation to other measures of speech perception. *Association for Research in Otolaryngology Abstracts*.

Kucera, F., & Francis, W. (1967). *Computational analysis of present day American English*. Providence, RI: Brown University Press.

Landauer, T. K., & Streeter, L. A. (1983). Structural differences between common and rare words: failure of equivalence assumptions for theories of word recognition. *Journal of Verbal Learning and Verbal Behavior, 12*, 119-131.

Lively, S. E., Pisoni, D. B., & Goldinger, S. D. (1994). Spoken word recognition: Research and theory. In M. Gernsbacker (Ed.), *Handbook of Linguistics* (pp. 265-301). New York: Academic Press.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Luce, P. A. (1986). *Neighborhoods of words in the mental lexicon*. (Research on Speech Perception Technical Report No. 7). Bloomington, IN: Speech Research Laboratory, Indiana University.

Luce, P. A., & Pisoni, D. B. (in press). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing*.

Luce, P. A., Pisoni, D. B., & Goldinger, S. D. (1990). Similarity neighborhoods of spoken words. In G. T. M. Altmann, (Ed.), *Cognitive Models of Speech Processing: Psycholinguistic and Computational Perspectives* (pp. 122-147). Cambridge, MA: MIT Press.

Meyer, T. A. & Bilger, R. C. (1997). Effect of set size and method on speech-reception thresholds in noise. *Ear and Hearing, 18*, 202-209.

Miller, G. A. (1946). Some characteristics of human speech. In S. S. Stevens, (Ed.), *Transmission and Reception of Sounds Under Combat Conditions* (pp. 58-68). Washington, DC: Office of Scientific Research and Development.

Miller, G. A., & Nicely, P. E. (1955). An analysis of perceptual confusions among some English consonants. *Journal of the Acoustical Society of America, 27*, 338-352.

Nusbaum, H. C., Pisoni, D. B., & Davis, C. K. (1984). Sizing up the Hoosier mental lexicon: Measuring the familiarity of 20,000 words. *Research on Speech Perception Progress Report No. 10*. Bloomington, IN: Speech Research Laboratory, Indiana University.

Penrod, J. P. (1994). Speech threshold and word recognition/discrimination testing. In J. Katz, (Ed.). *Handbook of Clinical Audiology* (pp. 147-164).

Pisoni, D. B., Nusbaum, H. C., Luce, P. A., & Slowiaczek, L. M. (1985). Speech perception, word recognition and the structure of the lexicon. *Speech Communication, 4,* 75-95.

Rabinowitz, W. M., Eddington, D. K., Delhorne, L. A., & Cuneo, P. A. (1992). Relations among different measures of speech reception in subjects using a cochlear implant. *Journal of the Acoustical Society of America, 92,* 1869-1881.

Sommers, M. S., Kirk, K. I., & Pisoni, D. B. (1997). Some considerations in evaluating spoken word recognition by normal-hearing and cochlear implant listeners I: The effects of response format. *Ear and Hearing, 18,* 89-99.

Stevens, S. S. (1951). Mathematics, measurement, and psychophysics. In S. S. Stevens, (Ed.), *Handbook of Experimental Psychology* (pp 1-49). New York: Wiley.

Tillman, T. W., Carhart, R., Wilber, L. (1963). A test for speech discrimination composed of CNC monosyllabic words. *Technical Documentary Report No. SAM-TDR-62-153.* USAF School of Aerospace Medicine, Brooks Air Force Base, Texas.

Treisman, M. (1978a). A theory of the identification of complex stimuli with an application to word recognition. *Psychological Review, 85*, 525-570.

Treisman, M. (1978b). Space or Lexicon? The word frequency effect and the error response frequency effect. *Journal of Verbal Learning and Verbal Behavior, 17*, 37-59.

Tyler, R. S., & Moore, B. C. (1992). Consonant recognition by some of the better cochlear-implant patients. *Journal of the Acoustical Society of America, 92*, 3068-3077.

Wang, M. D., & Bilger, R. C. (1973). Consonant confusions in noise: A study of perceptual features. *Journal of the Acoustical Society of America, 54,* 1248-1266.

Welkowitz, J., Ewen, R. B., & Cohen, J. (1991). *Introductory Statistics for the Behavioral Science.* Orlando, FL: Harcourt Brace Jovanovich.

Wilson, R. H., & Margolis, R. H. (1983). Measurements of auditory thresholds for speech stimuli. In D. F. Konkle & W. F. Rintelmann, (Eds.), *Principles of Speech Audiometry* (pp. 79-126). Baltimore: University Park Press.

Young, L. L., Dudley, B., & Gunter, M. B. (1982). Thresholds and psychometric functions of the individual spondaic words. *Journal of Speech and Hearing Research, 25*, 586-593.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 21 (1996-1997)
Indiana University

Effects of Alcohol on the Production of Words in Context: A First Report[1]

Steven B. Chin,[2] Nathan R. Large, and David B. Pisoni[2]

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405

---

[1] This research was supported by grants to Indiana University Bloomington from the Alcoholic Beverage Medical Research Foundation and the National Institutes of Health/National Institute on Deafness and Other Communication Disorders, Training Grant DC00012. We are grateful to Jon M. D'Haenens for his invaluable and generous assistance and technical advice on this project.

[2] Also DeVault Otologic Research Laboratory, Department of Otolaryngology–Head and Neck Surgery, Indiana University School of Medicine, Indianapolis, Indiana

S.B. CHIN, N.R. LARGE, AND D.B. PISONI

Effects of Alcohol on the Production of Words in Context: A First Report

Abstract. This study examined the effects of alcohol on acoustic parameters of spoken words in sentential contexts. Nine male talkers recorded 66 isolated sentences in a shadowing task in both a nonalcohol (BAC = 0.00%) and an alcohol condition (BAC ≥ 0.10%). From these sentences, 79 keywords were isolated and analyzed in terms of duration, fundamental frequency, probability of voicing, RMS amplitude, and AC peak. All talkers exhibited an increase in mean word duration; this increase was significant across talkers as well as for seven of the individual talkers. There was no significant difference in mean fundamental frequency, but the variability of fundamental frequency increased significantly from the nonalcohol to the alcohol condition. No significant differences were found between conditions in probability of voicing and AC peak. However, RMS amplitude increased significantly from the nonalcohol to the alcohol condition. These results are consistent with previous findings regarding the effects of alcohol on acoustic-phonetic parameters of speech.

Introduction

It is a commonplace observation that, among the various effects of alcohol consumption, the ability to produce speech is affected. Although superficial awareness of alcohol's effect on speech is millennia-old, the scientific study of these effects is not much older than this century (see Chin & Pisoni, 1997). During the twentieth century, the effects of alcohol on speech have been examined from a number of different perspectives, including medicine, psychology, and speech science.

Before the application of instrumental acoustic analysis to the investigation of alcohol effects on speech, research on this topic was generally conducted by medical and psychological investigators, and methodologies were generally derived from these disciplines. Especially insofar as speech is viewed as a behavior, early twentieth-century research drew on the expanding field of experimental psychology. Dodge and Benedict (1915), for instance, investigated voice reaction times to visual stimuli both with and without a standard dose of alcohol. Hollingworth (1923) measured the total time necessary for subjects to say aloud the names of colors of printed squares. Forney and Hughes (1961) were among the first to examine specifically the nature of speech produced under alcohol; they found that speech errors of various types (e.g., omissions, mispronunciations) did not increase significantly after alcohol consumption. Analysis of speech errors was a common methodology employed after Forney and Hughes (1961). Trojan and Kryspin-Exner (1968) observed increases in both syntactic and phonetic errors after consumption of alcohol, as well as changes in voice quality. In a study of Bulgarian-speaking talkers, Zaimov (1969) found that various types of errors (e.g., substitutions, omissions, repetitions) at various linguistic levels (e.g., word, syllable) increased as a function of increased alcohol consumption. Sobell and Sobell (1972) likewise found that errors committed while reading aloud a standard passage increased after alcohol consumption.

Alcohol's depressant effects on the central nervous system are manifested at both the cognitive and the motoric levels of functioning, although changes in speech production have most often been viewed as decrements in motor control. Furthermore, the changes that occur have for the most part been investigated from an analysis of the radiated speech signal, either acoustically or auditorally. A few researchers, however, have conducted more direct examinations of morphological or physiological changes in the organs of speech production that occur after consumption of alcohol. Investigation in this area has generally concentrated on the vocal folds. Dunker and Schlosshauer (1964) examined hoarseness in 46 German-speaking subjects using high-speech photography and stroboscopy. One of the subjects was instructed to "yell and sing with full voice when celebrating and consuming alcoholic beverages liberally" (p. 165). Although the effects of alcohol were most likely confounded with those of overusing the voice, Dunker and Schlosshauer concluded in any case that any effects were minimal. More recently, Künzel, Braun, and Eysholdt (1992) performed endoscopic and stroboscopic examination of the vocal folds. No morphological changes (specifically edema) were observed, although changes in vocal fold function were found: decreased vibration amplitude, reduced mucosal waves, and insufficient vocal fold adduction. Along with changes in fundamental frequency and intensity, Watanabe et al. (1994) also found vocal fold injections and edema on fiberscopic examination of Japanese talkers after alcohol. Whereas visual observation of the vocal folds is fairly observer-dependent and subjective, electroglottography is both a noninvasive and a relatively objective (although indirect) method of assessing laryngeal function. This method was employed by Johnson et al. (1993), who hypothesized that open quotient (i.e., the ratio of the glottal open phase to the entire glottal wave cycle) would decrease if there were morphological changes in the vocal folds (e.g., edema, swelling) resulting from alcohol consumption. In fact, they found that open quotient tended to increase as a function of rising blood-alcohol concentration (BAC).

Instrumental acoustic analysis of alcohol effects on speech is a relatively recent development, dating back only to the mid 1970s. From spectrographic analysis, Lester and Skousen (1974) found that sound substitutions and both consonant and vowel lengthening occurred in speech produced under alcohol. Fontan, Bouanna, Piquet, and Wgeux (1978) noted vowel formant compacting, vowel lengthening, voicing, frication, vocalization, and aspiration in the (French) speech of acutely alcoholized alcoholic subjects. Klingholz, Penning, and Liebhardt (1988), in a study of German-speaking talkers, found that frequency distributions of fundamental frequency, signal-to-noise ratio, and long-term average spectrum could discriminate speech produced with and without alcohol with a less than 5% error rate.

In a large-scale, comprehensive study of alcohol effects on the (German) speech of 33 acutely alcoholized subjects, Künzel, Braun, and Eysholdt (1992) found that alcohol brought about changes in nasality, segmental lengthening, incomplete articulations, changes in fundamental frequency, slowing in speaking rate, and increases in the number and duration of pauses. In a study of voice onset time (VOT), Swartz (1992) found that VOT variability appeared to be resistant to the influence of alcohol. Hollien and Martin (1996) reported that fundamental frequency increased for most of their subjects, and passage durations tended to increase, both as functions of increments in intoxication. Cummings, Chin, and Pisoni (1996) reported that pitch and intensity measures revealed no significant differences between speech produced with and without alcohol, but that jitter (frequency perturbation) and shimmer (amplitude perturbation) increased under alcohol. Additionally, using a glottal inverse filtering procedure to derive glottal excitation waveforms, Cummings et al. demonstrated that the glottal waveshape is less consistent and the vocal tract less stationary under alcohol.

The database used in the present study has been subjected to a number of analyses. Initially, Pisoni, Yuchtman, and Hathaway (1986) performed several perceptual, transcription, and acoustic analyses of speech data from four talkers. Perceptual results showed that both experienced (laboratory staff) and naive (college undergraduate) listeners could reliably discriminate sentences spoken with and without alcohol. Phonetic transcription by trained listeners revealed various types of errors in speech produced under alcohol, including vowel lengthening, deletions, partial articulations, and deaffrication. Acoustic analysis revealed, among other things, reliable changes in durational and timing aspects of speech produced under alcohol. Pisoni and Martin (1989) reiterated the perceptual findings of Pisoni et al. (1986), using Indiana State Police officers as listeners. Behne and Rivera (1990) examined alcohol effects on both segmental and prosodic properties of spondees. They found that, under alcohol, speech was produced with lower second and third formants, increased amplitude and amplitude variability, and increased fundamental frequency and fundamental frequency variability. Behne, Rivera, and Pisoni (1991) examined durational effects of alcohol in isolated words, sentences, and passages. They found that although isolated sentences and sentences within passages were reliably longer under alcohol, the durations of isolated monosyllabic and spondaic words were not reliably different between alcohol and nonalcohol conditions

In the present study, we examined the effects of alcohol on various acoustic parameters of single words in sentential contexts.

Method

The study of words in context reported here was part of a larger study conducted at the Speech Research Laboratory at Indiana University Bloomington under contract with General Motors Research Laboratories (Warren, MI).

Subjects

Subjects for this study were nine male volunteers at Indiana University between 21 and 26 years old who were recruited through a newspaper advertisement. They were paid to serve as subjects. All were native speakers of English, and none had a history of speech, hearing, or language disorder. All subjects completed a set of questionnaires to assess alcohol consumption and risk for alcoholism: (1) the short Michigan Alcoholism Screening Test (Selzer, Vinokur, & Van Rooijen, 1975), (2) the MacAndrew Scale (MacAndrew, 1965), (3) the socialization subscale of the California Psychological Inventory (Gough, 1969), and (4) a short alcohol-consumption questionnaire. Only subjects shown by these assessments to be moderate social drinkers at low risk for alcoholism were included in the study. Table 1 shows demographic data and screening test scores for the nine subjects.

Speech Samples

Subjects provided the following types of speech samples for analysis: 204 isolated monosyllabic words, 38 isolated spondaic words, 66 isolated simple sentences, 15 isolated alliterative sentences, and 3 connected passages. Monosyllabic words and the isolated simple sentences were presented auditorally in a shadowing task. The spondaic words and alliterative sentences were stored in digital files on a PDP 11/34 computer for visual presentation on a CRT video screen during recording. The three passages were typed separately on white paper for reading. The auditory stimuli (204 monosyllabic words, 66 simple sentences) were prerecorded in citation form by a male talker in a sound-attenuated booth using an Electro-Voice (Model D054) microphone and an Ampex AG-500 tape recorder. Stimuli were then low-pass filtered at 4.8 kHz and digitized at a 10 kHz sampling rate through a 12-bit A/D converter. A digital waveform editor (Luce & Carrell, 1981) was used in conjunction with a PDP 11/34 computer to edit the speech stimuli into separate digital files for later playback. Four audio tapes were produced using a computer-controlled audio tape-making program. The digital waveforms were output through a 12-bit D/A converter, low-pass filtered at 4.8 kHz and recorded on audio tape at a speech of 7.5 ips. On two of the audiotapes, the monosyllabic words were recorded first, with one second of silence after each word. The simple sentences were then recorded with three seconds of silence after each sentence. A different random ordering of words and sentences was used for each tape. The remaining two tapes were constructed similarly, but the sentences preceded the monosyllabic words.

Table 1: Subject Characteristics

Subject	Age	Initial BAC[a]	Final BAC[b]	MAST[c]	SOC[d]	MAC[e]	Alcohol intake, prior 30 days[f]
1	26	.10%	.10%	2	35	22	6.15
2	22	.16%	.10%	3	39	23	3.53
3	21	.17%	.10%	5	30	27	16.80
4	21	.13%	.075%	4	29	20	5.15
5	22	.15%	.085%	5	31	27	23.20
6	25	.135%	.15%	7	33	18	26.99
7	21	.15%	.10%	6	36	24	8.94
8	21	.15%	.095%	0	42	22	13.13
9	21	.19%	.12%	6	34	27	5.54

[a]Blood-alcohol concentration at beginning of recording session; [b]Blood alcohol concentration at end of recording session; [c]Short Michigan Alcoholism Screening Test (Selzer, Vinokur, & Van Rooijen, 1975); [d] Socialization subscale of California Psychological Inventory (Gough, 1969); [e]MacAndrew Scale (MacAndrew, 1965); [f]Self-reported alcohol intake during 30 days prior to experimental session (converted to fluid ounces 200-proof alcohol).

Speech materials for the study reported here consisted of the 66 isolated simple sentences, taken from Borden (1971), each containing one or two "keywords" on which acoustic measurements were made. The keywords from 34 of the sentences had been analyzed previously by Borden (Borden, 1971; Borden, Harris, & Oliver, 1973; Borden, Harris, & Catena, 1973) in studies of the effects of oral anesthesia (2% lidocaine) on speech; in addition, segmental and sentential acoustic analyses of these 38 keywords and 34 sentences had been performed by Pisoni, Yuchtman, and Hathaway (1986) in their study of alcohol effects on speech. One or two keywords from each of the remaining 32 sentences were selected for the present study, according to criteria established in Borden (1971) for the original 38 keywords, that is, with weighting toward fricatives and consonant clusters. Forty-one keywords were thus selected from the 32 sentences, so that analyses were performed on a total of 79 keywords.

Subject Preparation

Subjects participated in two recording sessions, one with alcohol ("alcohol") and one without ("nonalcohol"). Subjects agreed not to eat or drink for four hours prior to each experimental session. When subjects arrived for the nonalcohol session, blood-alcohol concentration (BAC) was measured from breath with a Smith & Wesson Breathalyzer (Model 900A) to insure an absence of alcohol in the system. On arrival for the alcohol session, subjects were administered a breath-alcohol test and then weighed before administration of alcohol. The alcohol preparation consisted of one part 80-proof (40%) vodka to three parts orange juice, administered in a dose of 1 gm alcohol per kg body weight, designed to raise BACs to 0.10 gm alcohol/100 ml blood (0.10%) over a 45-minute period. Subjects consumed a third of the total dose every 15 minutes. After 45 minutes, subjects rinsed their mouths and were administered a second breath-alcohol test. When BACs were at least 0.10%, audio recordings were made of subjects' speech samples.

Procedure

During recording sessions, subjects were seated in a sound-attenuated IAC booth, wearing a pair of matched and calibrated TDH-39 headphones with a boom-attached EV C090 L0-Z condenser microphone. Microphone placement was 4 inches directly in front of the subject's mouth. Auditory stimuli for shadowing were presented through the headphones from audiotape. Subjects were instructed to listen carefully to each stimulus sentence and then to repeat it back (i.e., to "shadow" it) aloud as soon as possible. Subjects were also informed that they would have a limited amount of time to repeat each sentence. All spoken responses were audio-recorded using a second Ampex AG-500 tape recorder. Recording levels were adjusted at the commencement of each subjects' initial recording session and, in order that amplitude could be compared across the two conditions, remained the same throughout both sessions.

All subjects remained naive to the actual purpose of the experiment. They were informed only that the experiment involved the effects of alcohol on memory and the rate at which they could shadow material after alcohol ingestion. No mention was made of the planned acoustic analyses that would be conducted, although this possibility was raised in the consent form that all subjects read and signed before the experiment began.

Analysis

All sentences were low-pass filtered (9.6 kHz cutoff) and digitized at a rate of 20,000 samples/sec. Each digitized sentence was stored in a computer file in ILS format (Interactive Laboratory Systems: Signal Technology, Inc.). To facilitate analysis in the present study, the digitized data files were then written to CD-ROM, originally in the same ILS format. However, because analysis was to be performed with software lacking native support for the ILS format, the files were transformed to the Entropic Waves SD format (Entropic Research Laboratory, Inc.) and written to a second CD-ROM (D'Haenens & Hernández S., 1995).

The digital speech files of productions of the sentences were then analyzed from the CD on a Sun SPARCstation 5 implementing ESPS/*waves+*, a digital signal processing and interactive display and editing software package (Entropic Research Laboratory, Inc., AT&T Bell Laboratories). Using time-aligned windows for waveforms, wide-band spectrograms, and labels, the onset and offset boundaries[3] of all 79 keywords in the 66 sentences were marked and labeled. A label consisted of the sample number and an alphanumeric marker to identify the labeled point. The labels thus created were saved in files and could be retrieved for use in subsequent acoustic analysis of signal properties between the marked boundaries.

Automated acoustic analysis of the keywords as segmented using the labeling program was implemented with a C-shell script incorporating the "get_f0" program, which calculated the following measures:

a. An estimation of fundamental frequency of the sampled-data speech file using the normalized cross-correlation function and dynamic programming (Secrest & Doddington, 1983; Talkin, 1995).

b. Probability of voicing

c. RMS amplitude

d. AC peak (peak amplitude of a wave above zero-current)

---

[3] In some cases, acoustic material from the preceding or the following word was included in the keyword. This occurred, for instance, when the keyword began with a sibilant and the preceding word ended with a sibilant (e.g., "hi*s sweater*"). In such cases, there was often continuous frication for the two sibilants, and the word boundary was ambivalent between the two words; therefore, the onset of the keyword was marked as the commencement of the entire span of frication. In all cases such as this one, however, uniform segmenting criteria were applied across conditions and talkers.

As mentioned, the label files contained a field indicating the temporal location of each label within the entire speech file. These timing indicators were imported into Microsoft Excel, where their differences were used to determine the total duration of each of the keywords.

Differences in durations were calculated by subtracting values in the nonalcohol condition from those in the alcohol condition. Differences in all measures obtained from "get_f0" were calculated by subtracting values in the alcohol condition from those in the nonalcohol condition. A series of paired *t* tests was performed to compare measurements from the nonalcohol and alcohol conditions, to determine effects of alcohol on words in context, for the following parameters:

a. Duration: Mean duration

b. Fundamental frequency (F0): Mean F0, Minimum F0, Maximum F0, Standard deviation of F0

c. Probability of voicing: Mean probability of voicing, Standard deviation of probability of voicing

d. RMS amplitude: Mean RMS amplitude, Maximum RMS amplitude, Standard deviation of RMS amplitude

e. AC peak: Mean AC peak, Minimum AC peak, Maximum AC peak, Standard deviation of AC peak

Results

Durations

Durations of keywords ranged from 97 ms to 826 ms in the nonalcohol condition and from 178 ms to 910 ms in the alcohol condition. Mean duration across all keywords and all talkers was 466 ms (*SD* = 111) in the nonalcohol condition and 496 ms (*SD* = 116) in the alcohol condition. A paired *t* test showed the difference in mean duration between the two conditions to be highly significant ($t = 11.376$, $p < .0001$). Figure 1 compares mean durations across keywords for each talker in the nonalcohol and alcohol conditions.

Insert Figure 1 about here

As Figure 1 indicates, all talkers exhibited increased mean durations from the nonalcohol to the alcohol condition. For all but two talkers (Talkers 3 and 5), these differences were statistically significant, as indicated in Table 2.

BEST COPY AVAILABLE

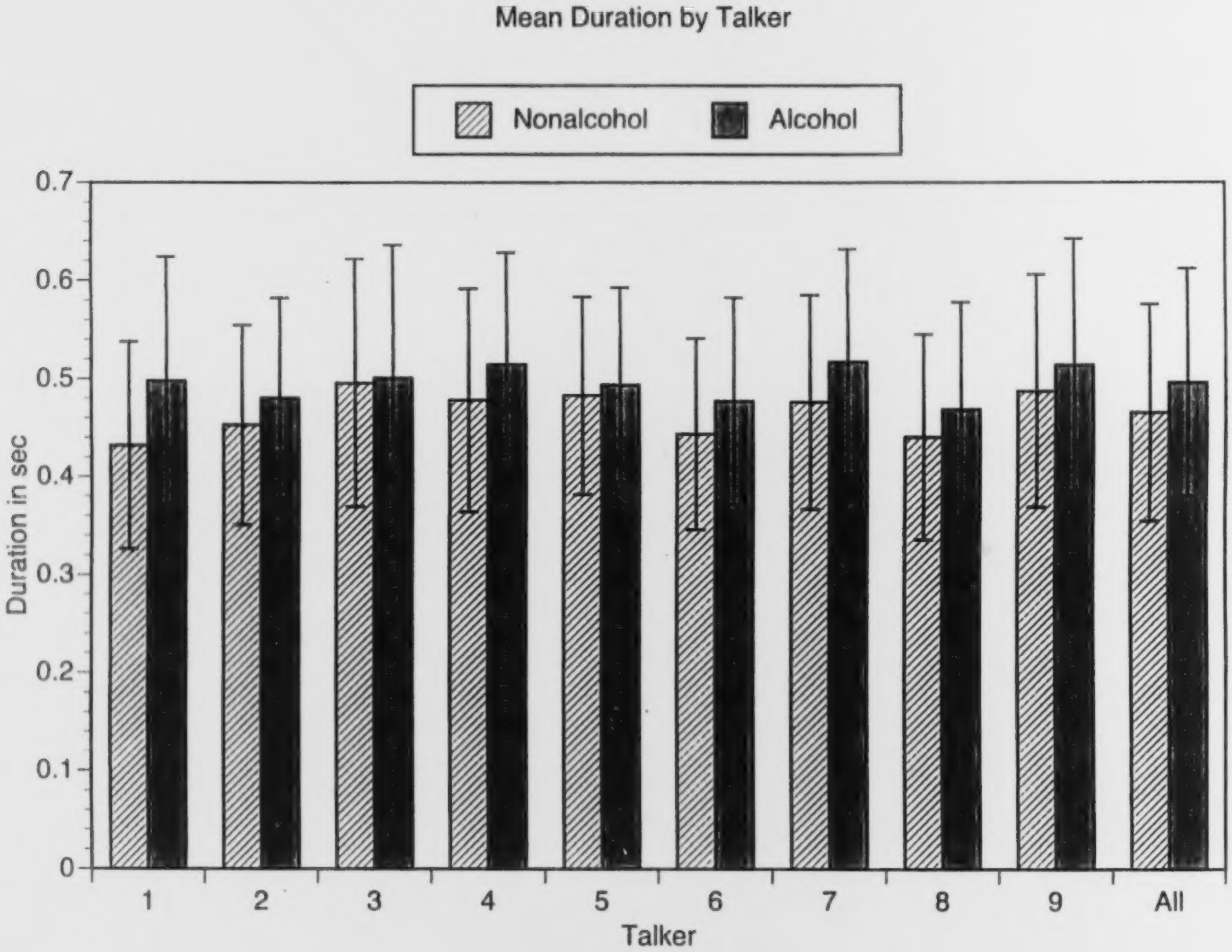


Figure 1: Mean duration (in seconds) of 79 sentence-embedded keywords by individual talker in nonalcohol (BAC = 0.00%; lighter bars) and alcohol (BAC ≥ 0.10%; darker bars) conditions. Error bars indicate standard deviations.

Table 2: Duration: Paired *t* test by individual talker

Talker	Mean Diff (ms)	DF	*t* value
1	66	78	7.978***
2	26	78	3.589**
3	8	78	.906
4	35	76	3.495**
5	10	78	1.455
6	35	78	5.073***
7	41	78	6.159***
8	28	78	3.864**
9	27	78	3.240*

*p < .01; **p < .001; ***p < .0001

Fundamental Frequency

Table 3 presents descriptive statistics for measurements of fundamental frequency in the keywords. Table 4 shows results from paired *t* tests for measurements involving fundamental frequency:

Table 3: Fundamental Frequency (F0) Descriptive Statistics

	Nonalcohol		Alcohol	
	M	*SD*	*M*	*SD*
Mean	95.508	34.013	96.617	36.938
SD	30.241	21.045	31.794	25.191
Minimum	0.000	0.000	0.000	0.000
Maximum	219.960	114.910	279.130	218.32

Table 4: Fundamental Frequency (F0)

	Mean Difference	*DF*	*t*
Mean F0	-1.129	708	-1.272
Minimum F0	2.833	708	2.047*
Maximum F0	-5.683	708	-3.105*
Standard Deviation of F0	-2.915	708	-3.550**

*p < .05; **p < .001

As Table 4 indicates, differences in minimum F0, maximum F0, and standard deviation of F0 between the nonalcohol and alcohol conditions were also statistically significant; the difference in mean F0 between conditions was not significant. Across all talkers, the mean change in F0 from the nonalcohol to the alcohol condition was an increase of 1.13 Hz ($t = -1.272$). Figure 2 shows that of the nine talkers, four (Talkers 1, 5, 6, 9) increased mean F0 in the alcohol condition, and five talkers (Talkers 2, 3, 4, 7, 8) decreased F0. Only three of the changes were statistically significant: the decreases for Talkers 3 ($t = 2.413$, $p < .05$) and 7 ($t = 2.685$, $p < .01$) and the increase for Talker 5 ($t = -3.573$, $p < .001$).

Insert Figure 2 about here

Thus, for two-thirds of the talkers in this study, fundamental frequency did not change significantly under alcohol, and for individual talkers, the significant results were equivocal as to the direction of change in F0 after consumption of alcohol.

The change in minimum F0 from the nonalcohol to the alcohol condition was a mean decrease of 2.83 Hz, a marginally significant difference ($t = 2.047$, $p < .05$). There was also a mean increase in maximum F0, however, of 5.68 Hz, a significant difference ($t = -3.105$, $p < .01$). Thus, although mean F0 remained unchanged, there were significant changes at both ends of the frequency range, that is, a decrease in mean minimum F0 and an increase in mean maximum F0, which served to expand the frequency range of F0 by approximately 5.15 Hz.

Concomitant with expansion of the frequency range of F0 was an increase in the variability of F0, as measured by the standard deviation. Across talkers, the mean increase in standard deviation was 2.92 Hz from the nonalcohol to the alcohol condition, a significant increase in variability ($t = -3.550$, $p < .001$). As shown in Figure 3, across keywords, the standard deviation of F0 increased for six talkers (Talkers 1, 4, 5, 6, 8, 9) and decreased for three talkers (Talkers 2, 3, 7). Only two of the differences were statistically significant, both of them increases: Talker 5 ($t = -3.445$, $p < .001$) and Talker 6 ($t = -3.028$, $p < .01$).

Insert Figure 3 about here

Other Acoustic Measures

Mean probability of voicing (ranging from 0 to 1) was .509 (*SD* = .464) in the nonalcohol condition and .504 (*SD* = .470) in the alcohol condition. This difference was not significant ($t = 1.27$, $p = .205$). Similarly, mean AC peak was .596 (*SD* = .335) in the nonalcohol condition and .592 (*SD* = .336) in the alcohol condition, also a nonsignificant difference ($t = 1.422$, $p = .156$).

Across keywords, mean RMS amplitude ranged from 129.96 to 1478.48 in the nonalcohol condition and from 111.60 to 2239.83 in the alcohol condition. Mean RMS amplitude across talkers was 481.44 (*SD* = 322.05) in the nonalcohol condition and 521.53 (*SD* = 359.48) in the alcohol condition, a significant difference ($t = -6.782$, $p < .0001$). Of the nine talkers in this study, seven showed increases in amplitude from the nonalcohol to the alcohol condition, and two showed decreases, as shown in Figure 4.

Insert Figure 4 about here

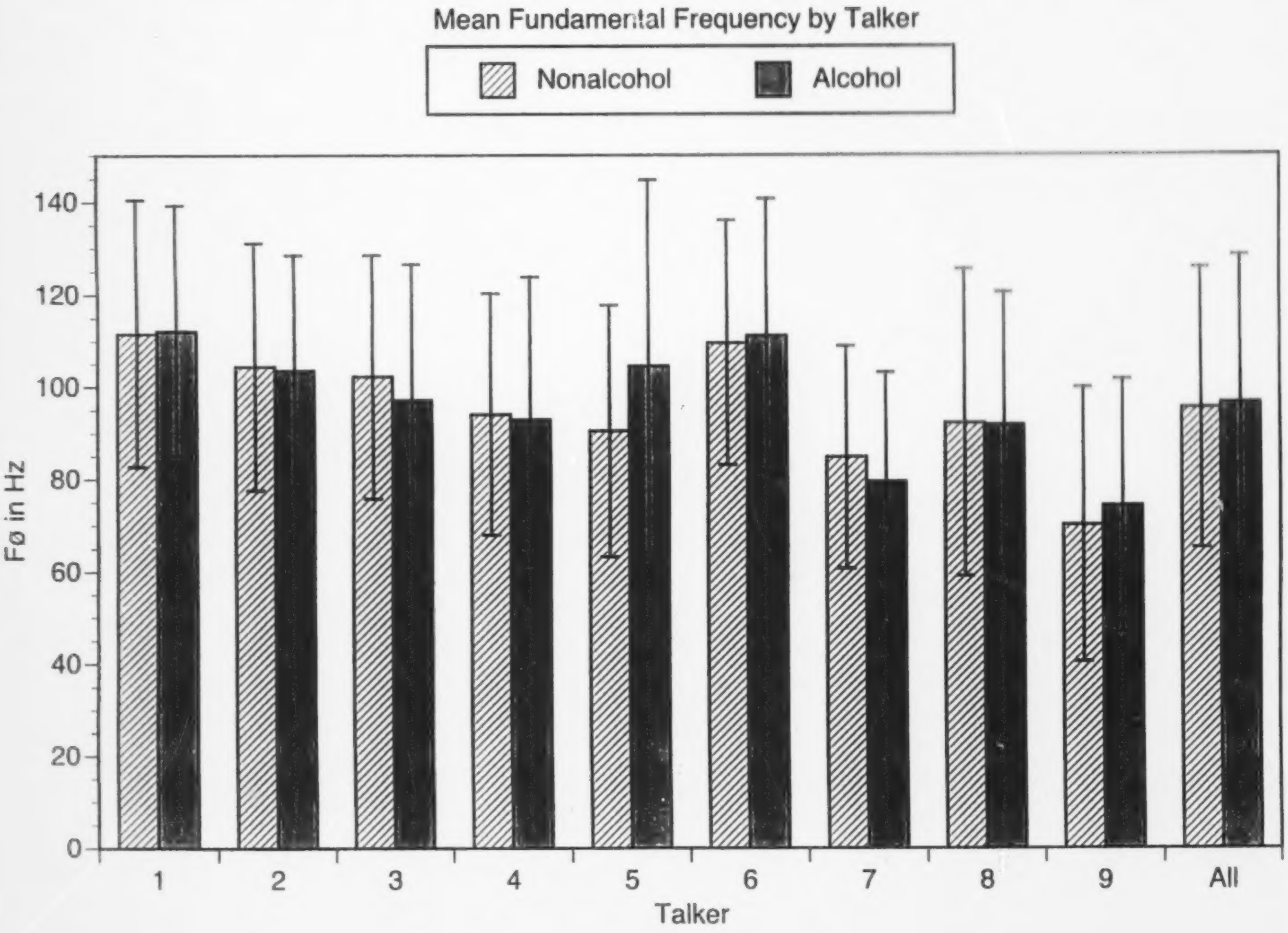


Figure 2: Mean fundamental frequency (in Hz) in 79 sentence-embedded keywords by individual talker in nonalcohol (BAC = 0.00%; lighter bars) and alcohol (BAC ≥ 0.10%; darker bars) conditions. Error bars indicate standard deviations.

BEST COPY AVAILABLE

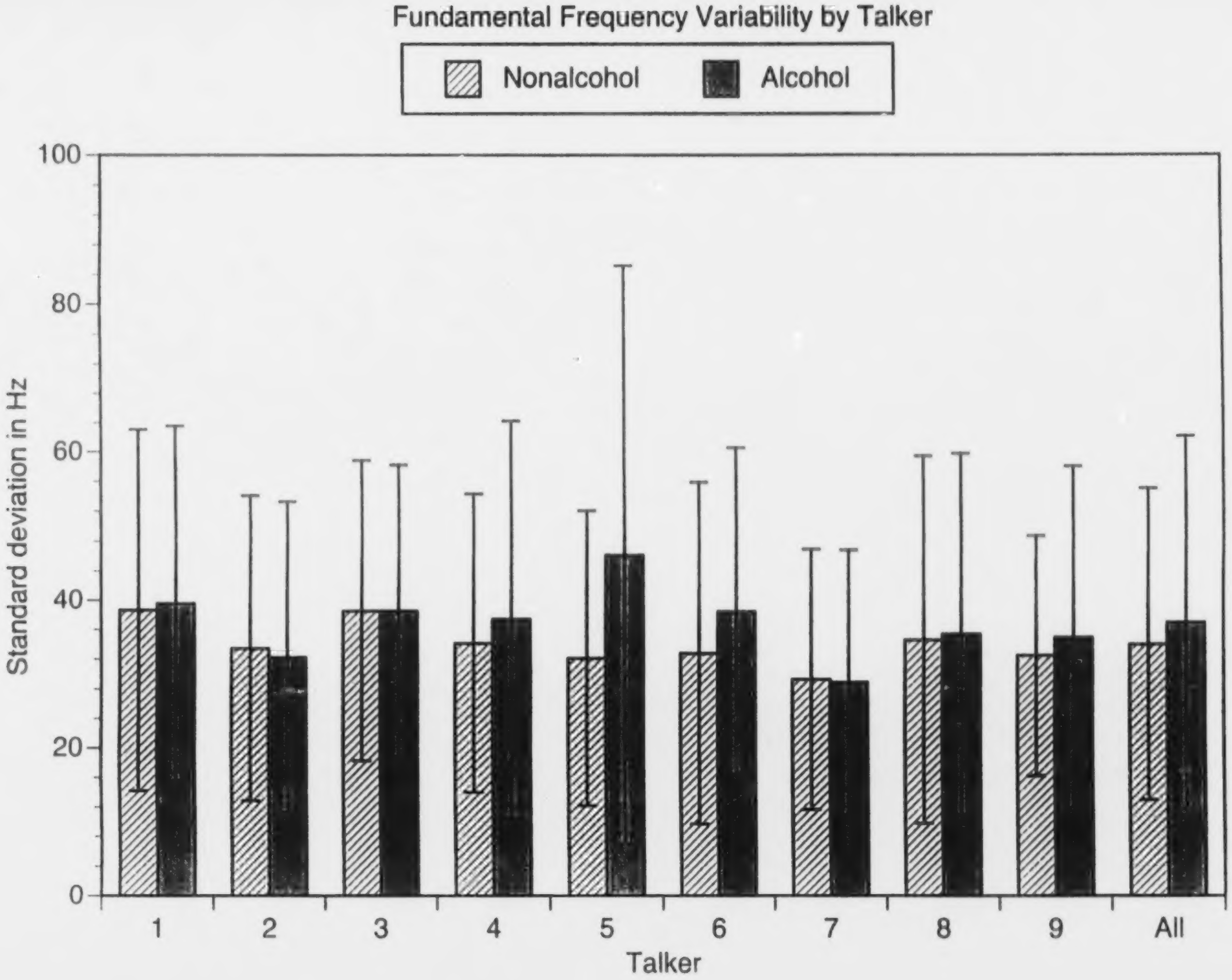


Figure 3: Mean standard deviation of fundamental frequency (in Hz) in 79 sentence-embedded keywords by individual talker in nonalcohol (BAC = 0.00%; lighter bars) and alcohol (BAC ≥ 0.10%; darker bars) conditions. Error bars indicate standard deviations.

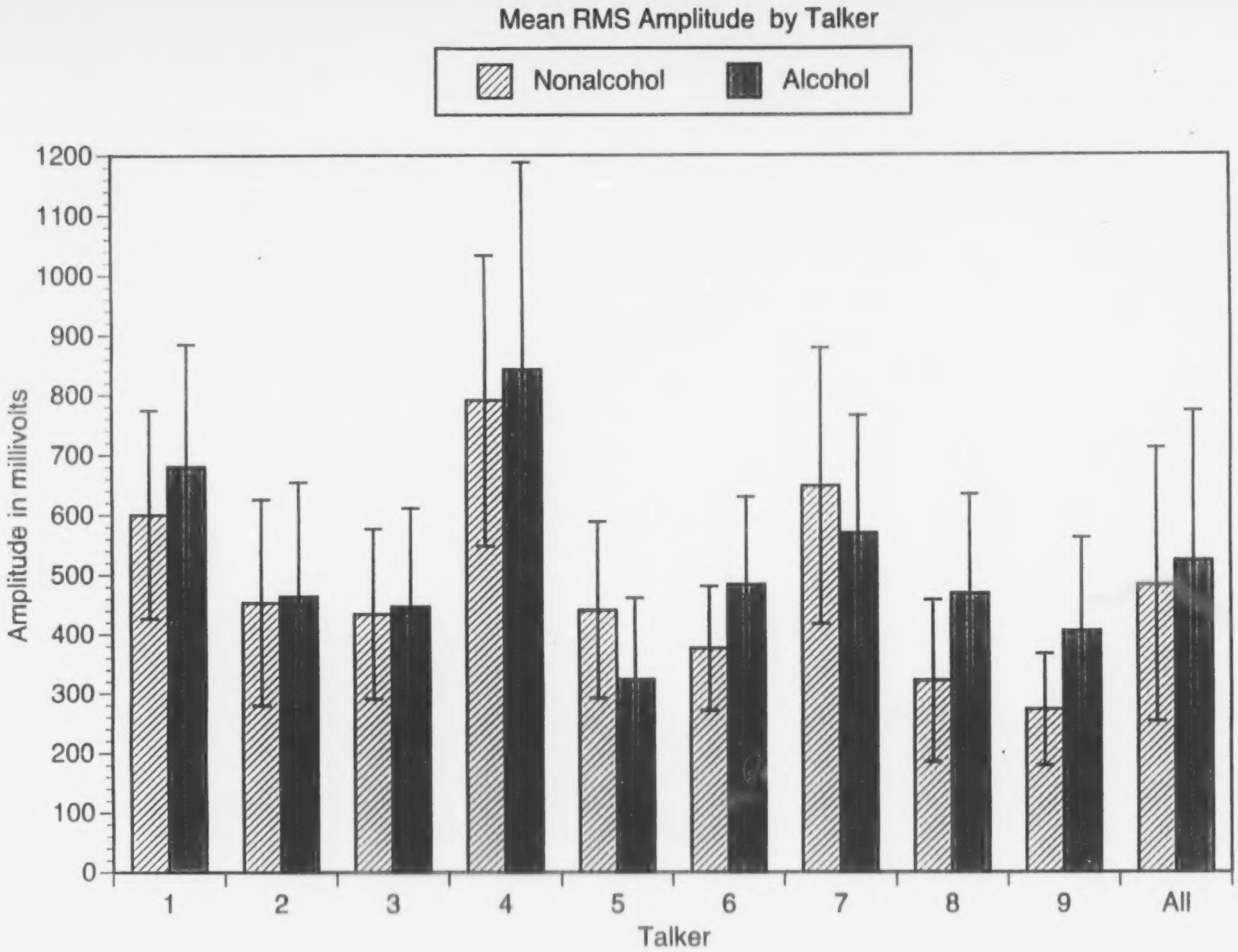


Figure 4: Mean RMS amplitude (in millivolts) in 79 sentence-embedded keywords by individual talker in nonalcohol (BAC = 0.00%; lighter bars) and alcohol (BAC ≥ 0.10%; darker bars) conditions. Error bars indicate standard deviations.

The increases in mean RMS amplitude were significant for Talkers 1, 4, 6, 8, and 9 ($p < .01$) but not significant for Talkers 2 and 3. Both cases of decreases in amplitude (Talkers 5 and 7) were significant ($p < .0001$). Mean maximum RMS amplitude across talkers was 1076.24 (*SD* = 481.44) in the nonalcohol condition and 1221.95 (*SD* = 249.27) in the alcohol condition; this increase was significant ($t = -8.90$, $p < .0001$).

Discussion

This study examined the effects of alcohol on the duration, fundamental frequency, and other acoustic parameters of words produced in sentential contexts. Consistent with a large number of previous studies, the most robust effect of alcohol in this study was on duration. Mean duration of keywords in the nonalcohol condition was 466 ms, and in the alcohol condition 496 ms, a significant increase ($p < .0001$). Previous studies have examined the effects of alcohol on the durations of a variety of linguistic units, including connected passages (e.g., Behne et al., 1991; DeJong, Hollien, Martin, & Alderman, 1995; Sobell et al., 1982), sentences (e.g., Behne et al., 1991; Pisoni et al. 1985), spondees (e.g., Behne et al., 1991), monosyllabic words (Behne et al., 1991), syllables (e.g., Behne & Rivera, 1990), consonants and vowels (e.g., Künzel et al., 1992; Pisoni et al., 1985, 1986), and VOT (e.g., Swartz, 1992). Across studies and across linguistic materials, durations were consistently and significantly greater under alcohol than without alcohol.

Particularly relevant in the present context is the earlier study by Behne et al. (1991). They examined duration effects of alcohol on isolated words, sentences, and passages, using the same speech materials from the same nine talkers as the present study. In the nonalcohol condition, mean durations of 34 of the sentences containing the keywords examined in the present study were 1510 ms in the nonalcohol condition and 1640 ms in the alcohol condition, a significant difference ($F_{1,30} = 182.67$, $p < .001$). Both monosyllabic words and spondees were also produced during the same sessions as the simple sentences. Mean durations of monosyllabic words were 521 ms in the nonalcohol condition and 540 ms in the alcohol condition, a significant increase ($F_{1,406} = 37.08$, $p < .001$). On the other hand, although mean durations of spondees were greater in the alcohol condition (529 ms) than in the nonalcohol condition (519 ms), this difference was not statistically significant ($F_{1,74} = 1.50$, n.s.).

In addition to measurements of duration, the literature on the effects of alcohol on speech also reports measures of fundamental frequency (F0), one of the acoustic correlates of perceived pitch. Both change in mean F0 and change in the variability of F0 have been examined. In the present study, F0 increased slightly but not significantly from the nonalcohol to the alcohol condition. Likewise, Künzel et al. (1992) reported a nonsignificant increase of 4.6 Hz. On the other hand, the increase in F0 was significant for vowels in spondees reported by Behne and Rivera (1990). Although these studies indicated a general trend toward increase in F0 under alcohol, individual talkers in all three cases could either increase or decrease F0. In the present study, four talkers increased F0 and five decreased it; moreover, only three of the changes were significant: two of them decreased and one of them increased. Behne and Rivera (1990) reported general increases for four of six talkers, a decrease for one talker, and both an increase and a decrease (depending on the vowel) for a sixth talker. Künzel et al. (1992) reported that of 33 talkers, F0 increased for 25, decreased for 7, and remained unchanged for one. Whereas the results from these studies indicate a general trend toward increase in F0 under alcohol, Watanabe et al. (1994) reported decreases in F0 for both male and female talkers (decreases of means from 115 to 108 Hz for males, and 262 to 246 Hz for females). Finally, Johnson, Pisoni, and Bernacki (1990) reported that F0 was lower closer to the time that their single subject (the captain of the U.S. Tankship *Exxon Valdez*) was alleged to have been intoxicated.

Although alcohol appears to exert no unique independent effect on mean F0, results reported in the literature regarding pitch variability are more uniform. Pitch variability is measured in the literature on alcohol and speech in two ways: (1) as the standard deviation of F0 and (2) as jitter, that is, peak-to-peak variations in F0. In the present study, the mean standard deviation of F0 across talkers increased significantly from the nonalcohol to the alcohol condition. Similar increases in standard deviation have been reported by Pisoni et al. (1985); Klingholz, Penning, and Liebhardt (1988); and Künzel et al. (1992). Similarly, increased jitter under alcohol was reported by Johnson et al. (1993), Künzel et al. (1992), Watanabe et al. (1994), and Cummings et al. (1996).

For the remaining acoustic parameters, alcohol appeared not to affect the probability of voicing for the keywords, indicating that words in sentential contexts were neither more nor less voiced under alcohol. Across talkers, AC peak was not affected by alcohol, but RMS amplitude was significantly higher in the alcohol condition. Increased RMS amplitude under alcohol was also reported by Johnson et al. (1993) for portions of sustained [a] and by Cummings et al. (1996) for the vowels in isolated words (using some of the same talkers as the present study).

The present study is the sixth from the General Motors/Indiana University database to appear in these progress reports (see Behne & Rivera, 1990; Behne et al. 1991; Cummings et al., 1995; Martin & Yuchtman, 1986; Pisoni et al., 1985). Results from these studies have shown that, for these nine talkers, the effects of alcohol (especially effects on duration) are consistent across various types of linguistic material, including isolated monosyllabic words, isolated spondaic words, sentences, and as demonstrated here, words produced in sentential contexts.

References

Behne, D. M., & Rivera, S. M. (1990). Effects of alcohol on speech: Acoustic analyses of Spondees. *Research on Speech Perception, Progress Report No. 16* (pp. 263-292). Bloomington, IN: Speech Research Laboratory, Indiana University.

Behne, D. M., Rivera, S. M., & Pisoni, D. B. (1991). Effects of alcohol on speech: Durations of isolated words, sentence, and passages. *Research on Speech Perception, Progress Report No. 17* (pp. 285-302). Bloomington, IN: Speech Research Laboratory, Indiana University.

Borden, G. J. (1971). *Some effects of oral anesthesia on speech: A perceptual and electromyographic analysis*. Ph.D. dissertation, City University of New York.

Borden, G. J., Harris, K. S., & Catena, L. (1973). Oral feedback II. An electromyographic study of speech under nerve-block anesthesia. *Journal of Phonetics*, **1**, 297-308.

Borden, G. H., Harris, K. S., & Oliver, W. (1973). Oral feedback I. Variability of the effect of nerve-block anesthesia upon speech. *Journal of Phonetics*, **1**, 289-295.

Chin, S. B., & Pisoni, D. B. (1997). *Alcohol and speech*. San Diego, CA: Academic Press.

Cummings, K. E., Chin, S. B., & Pisoni, D. B. (1995). Acoustic and glottal excitation analyses of sober vs. intoxicated speech: A first report. *Research on Spoken Language Processing, Progress Report No.20* (pp. 359-386). Bloomington, IN: Speech Research Laboratory, Indiana University.

Cummings, K. E., Chin, S. B., & Pisoni, D. B. (1996, May). Analysis of the glottal excitation of intoxicated versus sober speech: A first report. Paper presented at the 131st meeting of the Acoustical Society of America, Indianapolis, IN.

D'Haenens, J. M., & Hernández S., L. R. (1995). Using CD-ROM as a storage medium for digitized speech materials. In *Research on Spoken Language Processing, Progress Report No. 20* (pp. 403-407). Bloomington, IN: Speech Research Laboratory, Indiana University.

DeJong, G., Hollien, H., Martin, C., & Alderman, G. A. (1995). Speaking rate and alcohol intoxication [Abstract]. *Journal of the Acoustical Society of America*, **97**, 3364.

Dodge, R., & Benedict, F. G. (1915). *Psychological effects of alcohol: An experimental investigation of moderate doses of ethyl alcohol on a related group of neuro-muscular processes in man* (Publication No. 232). Washington, DC: Carnegie Institution of Washington.

Dunker, E. & Schlosshauer, B. (1964). Irregularities of the laryngeal vibratory pattern in healthy and hoarse persons. In D. W. Brewer (Ed.), *Research potentials in voice physiology (International Conference at Syracuse, 1961)* (pp. 151-184): [Syracuse:] State University of New York.

Fontan, M., Bouanna, J., Piquet, M., & Wgeux, F. (1978). Les troubles articulatoires chez l'éthylique. *Lille Médicale*, **23**, 529-542.

Forney, R. B., & Hughes, F. W. (1961). Delayed auditory feedback and ethanol: Effect on verbal and arithmetic performance. *Journal of Psychology*, **52**, 185-192.

Gough, H. G. (1969). *Manual for the California psychological inventory*. Palo Alto, CA: Consulting Psychologists Press.

Hollien, H., & Martin, C. A. (1996). Conducting research on the effects of intoxication on speech. *Forensic Linguistics*, **3**, 107-128.

Hollingworth, H. L. (1923). The influence of alcohol. *Journal of Abnormal Psychology and Social Psychology*, **18**, 204-237.

Johnson, K., Pisoni, D. B., & Bernacki, R. H. (1990). Do voice recordings reveal whether a person is intoxicated? A case study. *Phonetica*, **47**, 215-237.

Johnson, K., Southwood, M. H., Schmidt, A. M., Mouli, C. M., Holmes, A. T., Armstrong, A. A., Critz-Crosby, P., Sutphin, S. M., Crosby, R., McCutcheon, M. J., & Wilson, A. S. (1993). A physiological study of the effects of alcohol on speech and voice. Paper presented at the 22nd annual Symposium on the Care of the Professional Voice at the Voice Foundation.

Klingholz, F., Penning, R., & Liebhardt, E. (1988). Recognition of low-level alcohol intoxication from speech signal. *Journal of the Acoustical Society of America*, **84**, 929-935.

Künzel, H. J., Braun, A., & Eysholdt, U. (1992). *Einfluß von Alkohol auf Sprache und Stimme*. Heidelberg, Germany: Kriminalistik Verlag.

Lester, L., & Skousen, R. (1974). The phonology of drunkenness. In A. Bruck, R. Fox, & M. W. LaGaly (Eds.), *Papers from the Parasession on Natural Phonology* (pp. 233-239). Chicago, IL: Chicago Linguistic Society.

Luce, P. & Carrell, T. (1981). Creating and editing waveforms using WAVES. In *Research on Speech Perception, Progress Report No. 7* (pp. 287-297). Bloomington, IN: Speech Research Laboratory, Indiana University.

MacAndrew, C. (1965). The differentiation of male alcohol outpatients from nonalcoholic psychiatric outpatients by means of the MMPI. *Quarterly Journal of Studies on Alcohol*, **26**, 238-246.

Martin, C. S., & Yuchtman, M. (1986). Using speech as an index of alcohol intoxication. *Research on Speech Perception, Progress Report No. 12* (pp. 413-426). Bloomington, IN: Speech Research Laboratory, Indiana University.

Pisoni, D. B., Hathaway, S. M., & Yuchtman, M. (1985). Effects of alcohol on the acoustic-phonetic properties of speech. *Research on Speech Perception, Progress Report No. 11* (pp. 109-172). Bloomington, IN: Speech Research Laboratory, Indiana University.

Pisoni, D. B., & Martin, C. S. (1989). Effects of alcohol on the acoustic-phonetic properties of speech: Perceptual and acoustic analyses. *Alcoholism: Clinical and Experimental Research*, **13**, 577-587.

Pisoni, D. B., Yuchtman, M., & Hathaway, S. N. (1986). Effects of alcohol on the acoustic-phonetic properties of speech. In Society of Automotive Engineers (Ed.), *Alcohol, accidents, and injuries* (pp. 131-150). Warrendale, PA: Society Automotive Engineers.

Secrest, B. G., & Doddington, G. R. (1983). An integrated pitch tracking algorithm for speech systems. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing* (pp. 1352-1355). New York: Institute of Electrical and Electronics Engineers.

Selzer, M. L., Vinokur, A., & Van Rooijen, L. (1975). A self-administered short Michigan Alcoholism Screening Test (SMAST). *Journal of Studies on Alcohol*, **36**, 117-126.

Sobell, L. C., & Sobell, M. B. (1972). Effects of alcohol on the speech of alcoholics. *Journal of Speech and Hearing Research*, **15**, 861-868.

Sobell, L. C., Sobell, M. C., & Coleman, R. F. (1982). Alcohol-induced dysfluency in nonalcoholics. *Folia Phoniatrica*, **34**, 316-323.

Swartz, B. L. (1992). Resistance of voice onset time variability to intoxication. *Perceptual and Motor Skills*, **75**, 415-424.

Talkin, D. (1995). A robust algorithm for pitch tracking (RAPT). In W. B. Kleijn & K. K. Paliwal (Eds.), *Speech coding and synthesis* (pp. 495-518). Amsterdam: Elsevier Science.

Trojan, F., & Kryspin-Exner, K. (1968). The delay of articulation under the influence of alcohol and paraldehyde. *Folia Phoniatrica*, **20**, 217-238.

Watanabe, H., Shin, T., Matsuo, H., Okuno, F., Tsuju, T., Matsuoka, M., Fukaura, J., & Matsunaga, H. (1994). Studies on vocal fold injection and changes in pitch associated with alcohol intake. *Journal of Voice*, **8**, 340-346.

Zaimov, K. (1969). Die Sprachstörungen als Kriterium der Bewußtseinstrübungen. *Psychiatrie, Neurologie, und Medizinische Psychologie*, **21**, 218-225.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 21 (1996-1997)
Indiana University

Intelligibility of Normal Speech II: Analysis of Transcription Errors[1]

Amy T. Neel[2], Ann R. Bradlow[3] and David B. Pisoni

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405

---

[1] This work was supported by NIH-NIDCD Research Grant DC00111 to Indiana University, Bloomington, IN.

[2] Department of Speech & Hearing Sciences, Indiana University, Bloomington, IN.

[3] Department of Communication Sciences and Disorders, Northwestern University, Evanston, IL.

458

Intelligibility of Normal Speech II: Analysis of Transcription Errors

Abstract. Error analysis of a multi-talker database which contained transcriptions of 2000 sentences (20 talkers, 100 sentences) by normal-hearing listeners was used to investigate utterance- and talker-related factors affecting normal speech intelligibility. The 6100 transcription errors were classified into several categories and then analyzed to determine if any systematic patterns of phonetic, lexical, or talker characteristics affecting speech intelligibility emerged. Regarding phonetic factors, consonants were more susceptible to transcription error than vowels, and among consonants, non-sonorants, coronals, and word-final consonants were most vulnerable to error. Lexical status also affected error susceptibility: content words were more error-prone than function words. The patterns of error susceptibility for phonemes revealed in this study parallel the acoustic-phonetic characteristics of the speech signal that talkers spontaneously control when asked to speak clearly for hearing-impaired listeners suggesting that talkers have conscious or unconscious knowledge of the underlying acoustic-phonetic factors that affect speech intelligibility.

Introduction

Factors affecting speech intelligibility have been investigated for decades because of the importance of speech transmission in a number of fields. Several studies have consistently demonstrated differences in speech intelligibility among normal talkers (Black, 1957; Hood and Poole, 1980; Bond and Moore, 1994; Bradlow et al., 1996). In addition, speech intelligibility has been shown to differ substantially across a wide variety of speaking situations. For example, Picheny et al. (1985, 1986, 1989) showed that individual talkers displayed consistent acoustic-phonetic differences between "conversational" speech (speech directed to normal listeners) and "clear" speech (speech directed to hearing-impaired listeners). Individual talkers also display a range of intelligibility in connected speech. For instance, Pickett and Pollack (1963) found considerable within-talker variability in intelligibility scores for different samples excised from connected speech.

As part of an ongoing long-term investigation of factors that affect intelligibility of normal speech, this study focused on the patterns of listener errors obtained in a transcription task with sentences produced by normal talkers. By analyzing listener errors made in response to a large database of digitized sentences, we hoped to gain further insights into some of the utterance- and talker-related factors that affect normal speech intelligibility.

Some important factors related to speech intelligibility identified by previous studies include phonetic content, lexical characteristics, and talker-related acoustic-phonetic characteristics. Regarding phonetic content, the number and position of phonemes in the word as well as the specific phonemes affect intelligibility. Black (1952) found that some speech sounds enhanced intelligibility while others reduced intelligibility. Easily perceived sounds included vowels and diphthongs, voiced stops, affricates, and nasals. Sounds associated with reduced intelligibility included the voiceless obstruents /p/, /f/, and /›/. Black also found that words with many sounds were more intelligible than those with fewer sounds. In studying perceptual confusions of consonants in nonsense syllables, Miller and Nicely (1955) reported that voicing and nasality are transmitted more reliably in noise than other features. Perception of place of articulation was most affected by masking noise. House et al. (1965) found that initial consonants were perceived more successfully than final consonants.

Lexical characteristics such as word frequency or familiarity have also been shown to affect speech intelligibility. Black (1952) found that high frequency words, those with Thorndike ratings of 1 to 5, were more intelligible than low frequency words (Thorndike ratings from 9 to 10). He concluded that more familiar words are more accurately identified than less familiar words even among generally common words perhaps because the set size for familiar words is smaller and there are fewer opportunities for error. In studying the CID W-22 word lists given to hearing impaired listeners, Schultz (1964) found that less familiar words were more likely to be misidentified than more familiar words. He observed a marked tendency for highly familiar words to be substituted for incorrectly identified stimuli and suggested that highly familiar words, likely to be prominent among competing responses, are often chosen when stimuli are incorrectly perceived.

Hood and Poole (1980), however, found little evidence that unfamiliar words are more difficult to understand than familiar words. They selected 25 "difficult" and 25 "easy" words from a list of 500 words played to listeners (see Pisoni et al., 1985). When played to second group of listeners, these words retained a similar degree of difficulty. However, recordings of the same words produced by two additional speakers did not retain the same order when played to listeners - words which had previously been easy for subjects to perceive when produced by the first speaker became difficult when produced by the additional speakers. Hood and Poole concluded that although some of the difference in word difficulty is due to characteristics of the words themselves, the individual speakers play the dominant role in determining ordered difficulty of words. Luce (see Pisoni et al., 1985) also analyzed these 25 difficult and 25 easy words. Although the easy and difficult words did not differ significantly in frequency, the composition of their similarity neighborhoods differed. The neighborhoods of difficult words contained on average 56% of words of equal or greater frequency than the difficult words themselves, whereas easy word neighborhoods contained only 25% of words of equal or greater frequency. Therefore, difficult-to-perceive words had more "competition" from their neighbors than easy words.

Several studies have shown that there is a range of differences in speech intelligibility across normal speakers (Black, 1957; Hood and Poole, 1980; Bond and Moore, 1994). Bradlow et al. (1996) investigated acoustic-phonetic talker-specific characteristics that affected variability in speech intelligibility. They found that talkers who produced "precisely articulated" speech were generally more intelligible than other talkers who exhibited more reduction phenomena in their speech. For example, talkers who produced vowels widely dispersed in the phonetic vowel space had higher intelligibility scores than talkers with relatively reduced vowel spaces. Byrd (1994), in examining the TIMIT database of American English, found that male speakers exhibited more phonological reduction phenomena such as vowel centralization, flapping or alveolars, and fewer stop releases than female speakers. Differences in intelligibility among speakers may result from differences in amount of reduction phenomena in connected speech.

In the present study, we extended the earlier findings of Bradlow et al. (1996) by examining the transcription errors made by normal listeners in response to the same digitized speech samples that Bradlow et al. analyzed acoustically. Our error analysis addressed two specific questions: first, are some segments and words more likely to be misheard than others?, and second, do some talkers produce specific types of errors more than others? Errors in transcription were classified into several categories and entered into a database for analysis. Characteristics of the words and sentences in which errors occurred were then examined to determine which factors were related to speech intelligibility. The number and types of errors

BEST COPY AVAILABLE

produced by each talker were also examined to ascertain what speaker characteristics might be related to speech intelligibility.

Procedures

The sentence transcriptions used in the error analysis came from the Indiana Multi-Talker Sentence Database (Karl and Pisoni, 1994). The database consists of 100 Harvard sentences (IEEE, 1969) which are all single-clause sentences containing five key words plus some additional function words. Each of the 100 sentences was spoken by 20 talkers, 10 males and 10 females. None of the talkers had any known speech or hearing impairment. During the recordings, the sentence productions were monitored for any obvious misarticulations or disfluencies. The talkers read the sentences from a computer screen in a sound-attenuated booth. They spoke into a Shure (SM98) microphone, and the speech was digitized on-line using a 16 bit analog-to-digital converter (DSC Model 240) at a 20 kHZ sampling rate. Using a signal processing software package (Luce and Carrell, 1981), the average root mean square amplitude of each digital speech file was equated. The speech files were then converted to 12-bit resolution for presentation to listeners.

Sentence transcriptions were obtained from ten listeners for each talker for a total of 200 listeners. Subjects were Indiana University students, all native speakers of American English with no known speech or hearing impairment. Each group of ten listeners heard the list of the 100 Harvard sentences produced by one of the 20 talkers. The sentences were low-passed filtered at 10 kHz and presented binaurally through matched and calibrated TDH-39 headphones using a 12-bit digital-to-analog converter at a comfortable listening level (75 dB SPL). A PDP-11/34 computer was used to control the presentation of the stimuli and to record responses. Listeners typed what they heard on a computer keyboard.

Transcription errors were analyzed by number and type of errors for each word, each sentence, and each talker. Errors were classified into several categories and entered into a database in order to identify general patterns of errors. *Phonetic* errors were defined as transcription errors made on one or more phonetic segments within a word. Phonetic errors could be further classified as *consonant* errors, in which one or more consonants were incorrectly transcribed (for example "blew" was transcribed for "glue", *vowel* errors, in which one or more vowels in the word were incorrectly transcribed ("study" was substituted for "steady", or *consonant+vowel* errors, in which consonant and vowel errors occurred in the same word ("keep" for "kick"). Some phonetic errors could also be classified as semantic or typing/spelling errors (for example, "soft breeze" transcribed as "salt breeze". These were labeled as *phonetic/other* errors. *Typing/spelling* errors were defined as words with incorrectly typed letters (close in keyboard position to the intended key such as "blus" for "blue" or misspelled words ("throun" for "thrown". *Semantic* errors consisted of the substitution of another word for the word produced by the talker or switching of the order of the words produced by the talker. Semantic errors were always real words and were often related in meaning to the intended word (such as the substitution of "flame" for "fire". *Omissions* occurred when the word produced by the speaker was left blank by the listener in the transcription. *Additions* occurred when an extra word not produced by the talker was typed by the listener. Errors were classified as *unknown* when some letters were typed by the listener, but the error category could not be determined ("kn" for "gnawed").

Results and Discussion

Overall Error Distribution

Figure 1 shows the distribution of transcription errors made by all listeners for all 20 talkers. Typing and spelling errors were the most frequent type of transcription error, accounting for roughly one-

461

third of the 6100 total errors. Phonetic errors (errors on one or more consonants or vowels in a word) made up about one-quarter of the transcription errors, and omitted words made up nearly one-fifth of all errors made by listeners. Semantic errors were relatively infrequent (less than 10% of all errors) as were addition errors, phonetic/other errors and errors which could not be classified (each less than 3% of all errors).

Insert Figure 1 about here.

Phonetic Characteristics

Phonetic errors were examined more closely to determine if all phonetic segments are equally susceptible to transcription error. Consonant errors made up about 16% of all errors in the database while vowels comprised only about 5% of all errors. Errors on both consonant and vowel segments in the same word also made up about 5% of the total transcription errors. Chi-square tests were used to compare the distribution of consonant and vowel errors to the distribution of consonant and vowel segments in the 100 test sentences. This analysis revealed that consonants were more likely to be incorrectly transcribed than were vowels. Nearly 64% of the segments in the 100 sentences were consonants, but 77% of phonetic errors occurred on consonants ($P^2 = 106.5$, $p < .001$). The majority of words in the 100 sentences began with consonant phonemes (83%). There were somewhat more errors than expected for words beginning with consonants, whereas vowel-initial words had fewer errors than expected ($P^2 = 60.48$, $p < .001$).

Further analysis of the phonetic errors revealed that some consonants were more likely to be transcribed incorrectly by listeners than others. Chi-square tests were performed on consonant errors according to manner of articulation, place of articulation, sonorance, voicing, and position of the segment within the word. When results of the chi-square test indicated that the distribution of observed errors differed from that of the expected errors, the $(O-E)^2/E$ terms were examined to determine which types of segments were more often in error than expected and which segment types appeared resistant to error (see Table I). Nonsonorants (such as /s/ and /t/) were more likely to produce errors, and sonorants (such as /r/ and /m/) were less likely to produce errors than would be expected from the distribution ($P^2 = 222.7$, df = 1, $p < .001$). However, there were no significant deviations from expectations for errors on voiced and voiceless consonants ($P^2 = 1.80$, df = 1, $p < .178$). Consonants occurring in clusters were more susceptible to transcription error than singleton consonants ($P^2 = 112.5$, $p < .001$). Also, word-final consonants were more susceptible to transcription error while word-initial and -medial consonants appeared somewhat more resistant to error ($P^2 = 170.0$, df = 2, $p < .001$). In comparing consonants for manner of articulation, stops were more likely to be misperceived while nasals, liquids, and glides were less likely to be misperceived than would be expected based upon the distribution of consonant segments in the sentences ($P^2 = 247.8$, df = 5, $p < .001$). Regarding place of articulation, there were more errors on interdentals, alveolars, and palatals and fewer errors on velars, bilabials, and labiodentals than expected ($P^2 = 200.7$, df = 6, $p < .001$). From this analysis, it appears that a segment vulnerable to transcription error is a non-sonorant coronal or interdental consonant in a final cluster (e.g., the final /s/ in "costs". An error-resistant segment is a sonorant, non-coronal singleton in the initial position of a word (e.g., the /m/ in "many").

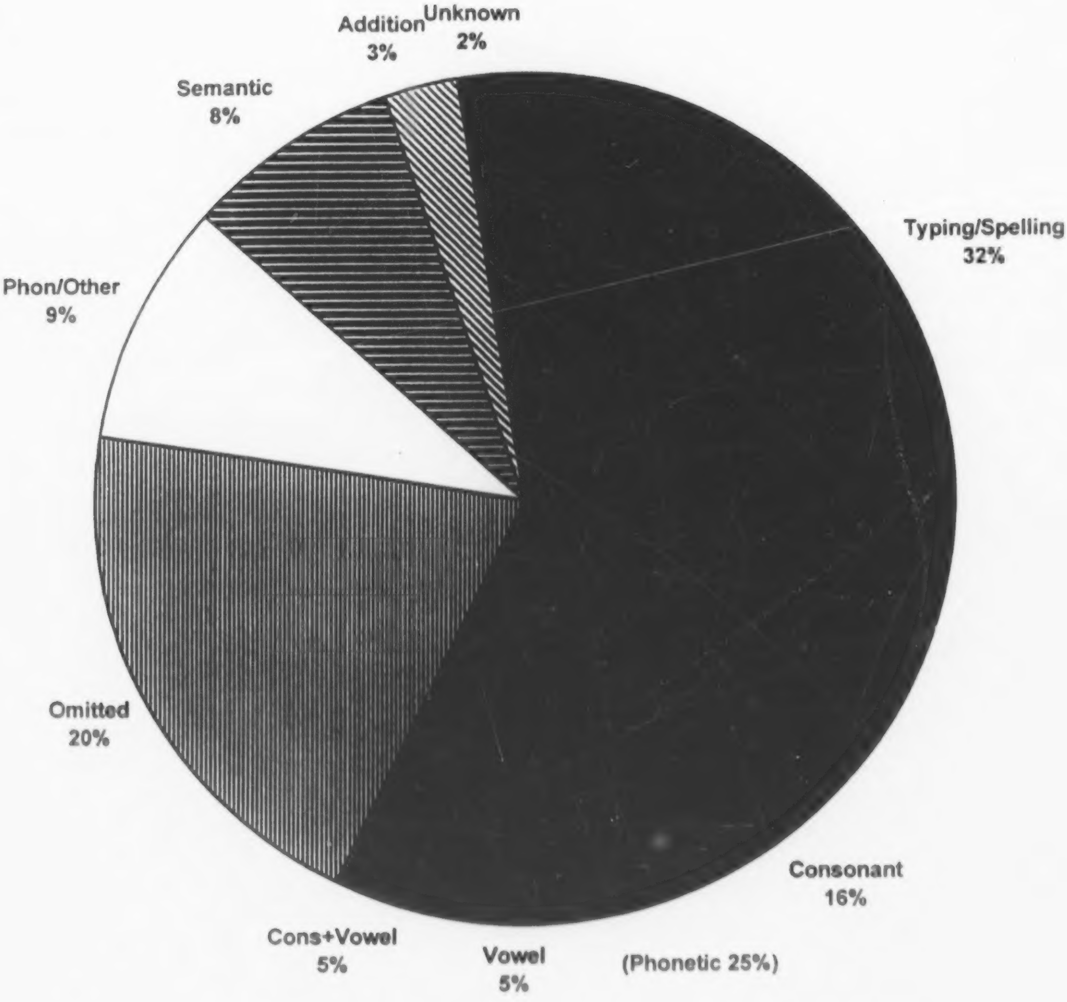

Addition
3%
Unknown
2%
Semantic
8%
Typing/Spelling
32%
Phon/Other
9%
Omitted
20%
Consonant
16%
Cons+Vowel
5%
Vowel
5%
(Phonetic 25%)

Table 1.

Analysis of consonant errors. For each characteristic, the number of occurrences in the database and the number of consonant errors are shown. Also listed are the terms included in the chi-square summary. Those features marked by an asterisk have fewer errors than expected based upon distribution in the database.

Manner	$\chi^2 = 247.44$, df = 5		$p < 0.001$
	Occurrences	**Errors**	**$(O - E)^2/E$**
Stop	521	98	43.95
Fricative	515	350	14.37
Affricate	24	39	29.34
Nasal	199	18	*76.26
Liquid	180	36	*41.74
Glide	74	0	*42.10

Position	$\chi^2 = 169.99$, df = 2		$p < 0.001$
	Occurrences	**Errors**	**$(O - E)^2/E$**
Initial	751	314	* 26.97
Medial	135	10	* 57.02
Final	627	517	86.00

Sonorance	$\chi^2 = 222.65$, df = 1		$p < 0.001$
	Occurrences	**Errors**	**$(O - E)^2/E$**
Sonorant	453	54	* 155.86
Nonsonorant	1060	787	66.80

464

Place	$\chi^2 = 200.68$, df = 6		$p < 0.001$
	Occurrences	**Errors**	**$(O - E)^2/E$**
Bilabial	213	62	* 26.87
Labiodental	96	24	* 16.19
Interdental	145	163	84.24
Alveolar	808	505	6.96
Palatal	46	47	17.89
Velar	173	31	* 44.19
Glottal	32	9	* 4.35

Number	$\chi^2 = 112.49$, df = 1		$p < 0.001$
	Occurrences	**Errors**	**$(O - E)^2/E$**
Singleton	979	400	* 39.37
Cluster	534	441	73.13

Voicing	$\chi^2 = 1.81$, df = 1		$p < 0.179$
	Occurrences	**Errors**	**$(O - E)^2/E$**
Voiced	627	866	* 0.74
Voiceless	364	474	1.07

Lexical Characteristics

The transcription errors were next examined to determine if any characteristics of the words affected susceptibility to error. The lexical characteristics included syntactic category (adjective, adverb, article, conjunction, noun, preposition, pronoun, or verb); the number of phonemes comprising the word; and lexical status indexed by frequency and familiarity. The 459 words in the database were also classified as content words (noun, verb, adjective, or adverb) or function words (article, preposition, or pronoun). About 60% of the words in the 100 sentences were content words: words that carry the meaning in sentences and can be morphologically complex. The remainder of the words were function words, words which play a grammatical role in sentences but do not carry semantic meaning. Errors occurred more often than expected on content words (about 71% of errors occurred on content words) and less often than expected for function words (P^2 = 247.5, df = 1, p < .001). (The content-function distinction will be discussed further in a later section.)

Nouns were the most common syntactic type (comprising about 30% of all words in the database), followed by articles (21%), and verbs (19%). Fewer errors than expected occurred on words classified as articles which are generally short in length and frequent in the lexicon, and more errors than anticipated occurred on verbs (P^2 =296.4, df = 7, p < .001). Verbs may be more susceptible to error due to the presence of word-final morphological endings which makes them less predictable than other syntactic types. When the category of semantic errors alone was examined, we found that transcription errors occurred more often than expected on adjectives and prepositions while articles appeared fairly resistant to semantic error. Adjectives appear particularly susceptible to substitution of similar or related words (semantic error) such as the substitution of "slick" for "smooth" in the sentence, "The birch canoe slid over the smooth planks," or the reversal of "fire" and "flame" in the sentence, "Smokey fires lack flame and heat." This may be due to the fact that adjectives have more synonyms than other syntactic types.

Word length also had some effect upon number of transcription errors as well. There were fewer errors than expected for words of one to two phonemes in length. Words that were four to six phonemes in length appeared somewhat more susceptible to error than expected. There were only a few seven to eight phoneme words in the database, so it was difficult to ascertain whether the longest words were particularly susceptible to error. However, the correlation of word length with error susceptibility scores was not significant (r = +.0619, p = .186).

In order to examine the distribution of errors according to lexical characteristics, word frequencies based on the Brown Corpus of printed text (Kucera and Francis, 1967) were obtained for each word in the database. In addition, an online version of Webster's Pocket Dictionary containing 20,000 entries was used to obtain familiarity ratings (Nusbaum et al., 1984) and neighborhood characteristics (Luce, 1986) for most of the words in the database.

The "similarity neighborhood" of a word is defined as the set of words differing from that word by one phoneme (Luce, 1986). Neighbors were computed by substituting, adding, or deleting one phoneme in any position of the word. For the word "gas", some of its 15 neighbors include the words "bass," "mass," "pass," "gab," "gal," and "guess." The density of a neighborhood is equal to the number of neighbors that can be generated, and the mean frequency of a neighborhood is the average word frequency of all the neighborhood members (Luce, 1986). For example, the frequency for the word "gas" is 98.0, and the mean frequency of its neighbors is 21.87. The prominence of a word in its neighborhood, how much it "stands out" among its neighbors, may be calculated by subtracting the mean neighborhood frequency from the frequency of that word. For example, the prominence of "gas" relative to its neighborhood is 76.13. Several studies have shown that words from low density neighborhoods (those that contain few neighbors) and words with higher frequencies than the mean frequency of their neighbors are more accurately and more quickly identified than words from high density neighborhoods or those which are not "prominent" in their neighborhoods (Pisoni et al., 1985; Luce, 1986; Luce et al., 1990). Familiarity, neighborhood density, mean neighborhood frequency, and prominence (word frequency - mean neighborhood frequency) were obtained for the 373 words (out of the total of 459 words in the sentence database) that were included in the online Webster=s Pocket Dictionary. Each of these five measures was then correlated with the number of errors on words in the database to determine if the frequency and neighborhood characteristics of a word influenced the likelihood of a transcription error.

An "error susceptibility score" based upon the chi-square statistic was calculated for each of the 459 words in the database. The formula to calculate the error susceptibility score used the observed and expected number of errors for each word:

(Observed number of errors - Expected number of errors)^2*/Expected number of errors.*

A negative sign was appended to the score if the observed number of errors was less than the expected number of errors. The "expected number of errors" term takes into account the fact that some words in the database occurred more than once in the 100 sentences and would be expected to have more errors based solely upon their frequency of occurrence than words which appeared only one time. For example, the word "and" appeared 14 times in the 100 sentences and comprised 1.8% of all the words in the set of sentences. If all the words were equally prone to error, then 1.8% of all 6120 errors should occur on the word "and." Thus, the expected number of errors for "and" is 110. The number of errors actually observed for the word "and," however, totaled only 92. Accordingly, an error susceptibility score of -3.07 was obtained for the word "and." The word "beat" occurs only once in the set of sentences. Therefore, only 0.13% of all errors (7.9 errors) was expected to occur for the word "beat" if all words were equally susceptible to transcription error. However, 48 errors were observed for "beat" yielding an error susceptibility score of 204 for this word. According to the error susceptibility scores, the word "and" is slightly resistant to transcription error, whereas "beat" is highly error-prone.

A low but significant negative correlation of word frequency with error susceptibility was obtained ($r = -.152$, $p < .001$). As previous studies have found (Black, 1952; Schultz, 1964), vulnerability to error increased as word frequency decreased. For the 373 words which were available in the online dictionary, the correlation of error susceptibility scores with familiarity was not significant ($r = -.082$, $p = .115$). There was little variation in familiarity scores for words in the Harvard sentences, so the lack of a significant correlation is not surprising. The correlation of error susceptibility scores with neighborhood density was also not significant ($r = -.044$, $p = .402$). Thus, the number of neighbors in the word's neighborhood did not appear to affect the likelihood of transcription error of words in sentences. Error susceptibility scores were significantly negatively correlated with the mean neighborhood frequency ($r = -.1051$, $p = .003$) and with the prominence of the word in its neighborhood ($r = -.1954$, $p = .000$): susceptibility to error increased as the frequency of a word compared to its neighbors decreased. All the significant correlations observed were very low, suggesting that word frequency and prominence of a word among its lexical neighbors play a relatively small role in controlling word intelligibility in sentence contexts.

Because the analyses described above revealed that several factors contribute to intelligibility of words in sentences, the words which were least susceptible to error and most susceptible to error (according to the error susceptibility scores) were examined in greater detail. Twelve of the 35 words least prone to error were function words, whereas 26 of the 28 words most vulnerable to error were content words. T-tests comparing the least and most susceptible words revealed that the least susceptible words were significantly shorter than the most susceptible words ($t = -3.408$, $df = 61$, $p < .002$). The least susceptible words were also significantly higher in frequency than the words most susceptible to error ($t = 2.439$, $df = 61$, $p < .018$). Furthermore, the least susceptible words had higher familiarity ratings than the most susceptible words ($t = 2.15$, $df = 54$, $p < .036$) and were more prominent in their neighborhoods than error-prone words ($t = 2.361$, $df = 54$, $p < .019$). Although lexical characteristics such as word length, frequency, familiarity, and prominence are not strongly related to error susceptibility for the database as a whole, they appear to be significant factors in distinguishing between the words least and most prone to error.

Several words which had substantially more errors than expected were examined to determine the possible causes for their reduced intelligibility. The word "walled" from the sentence "The walled town was seized without a fight," had many consonant errors: the final /d/ was omitted by over 50 listeners. Segmentation of the homorganic consonants /d/ and /t/ in the phrase "walled town" was problematic for listeners on 17 of the 20 speakers in the database. Listeners also added and deleted phonemes from several nouns and verbs. The word "beat" was transcribed as "beats" in the sentence "The heart beats strongly and with firm strokes," by 49 listeners spread across 13 of the speakers. Similarly, the word "lay" was heard as "laid" by many listeners in the sentence "The slush lay deep along the street." For 11 speakers, more than 30 listeners transformed "play" into "place" in the sentence AThe play seems dull and quite stupid." For the word "costs," nearly 40 consonant errors were counted with the final /s/ of the consonant cluster being deleted in the sentence "A pound of sugar costs more than eggs." Difficulty in segmenting boundaries between words and problems in detecting low-intensity final consonants, particularly in clusters, resulted in a number of transcription errors across many speakers. Bond and Garnes (1980) also noted a substantial number of segmental errors involving word boundaries in their corpus of 1000 misperceptions of fluent speech.

Although vowel errors were relatively rare, listeners had consistent difficulty with vowels in a few words. Forty-five listeners interpreted the word "steady" as "study" in the sentence "Four hours of steady work faced us." The word "pins" was transcribed as "pens" on 38 occasions. Also, the word "bail" was transcribed as "bill," "bell", or "build" by many listeners in the sentence "Bail the boat to stop it from sinking." These vowel substitutions tended to consist of vowels which were similar to the intended vowel. Semantic integrity tended to be preserved when vowel errors like these occurred.

A few words appeared to generate semantic errors. For instance, "sprained" and "stained" were substituted for "strained" in the sentence "The wrist was badly strained and hung limp," by nearly 30 listeners. "Salt" was transcribed as "soft" by 31 listeners in the sentence "The salt breeze came across from the sea." Several words, including "slick," "huge," and "small" were substituted for the intended word "smooth" in "The birch canoe slid over the smooth planks." Many semantic confusions were also phonetically related to the intended word produced by the speakers. As noted above, adjectives seem to be more prone to semantic errors than other syntactic types.

Finally, several words were susceptible to typing and spelling errors. Listeners had particular difficulty producing the words "cue," "gnawed," "hoist," "into," "lose," "squirrel," and "threw" using the computer keyboard.

Content-function Distinction

Differences in the pattern of errors between content and function words were noted throughout the analysis of the lexical characteristics of the transcription data. Content words were more susceptible to error than would be expected based upon the number of content words that appeared in the sentences whereas function words were less likely to produce error than expected. The content-function distinction was also important in separating the words least susceptible and most susceptible to error. Of the ten words least susceptible to transcription error, eight were function words. However, nine of the ten words most vulnerable to error were content words.

A comparison of content and function words in the database revealed several differences in their lexical characteristics. Content words had an average of 4.56 phonemes and function words only 3.11

phonemes, although this difference did not reach statistical significance ($t = -.8629$, $df = 457$, $p = .409$). Function words were much higher in frequency than content words (mean content frequency = 265.61, mean function frequency = 6462.83; $t = 10.06$, $df = 457$, $p < .001$), and function words were significantly more prominent in their neighborhoods than content words ($t = 8.68$, $df = 373$, $p < .001$). Thus, it is difficult to dissociate the effects of word frequency and syntactic status (content or function word) on error susceptibility. Some words could be less error-prone than others simply because they are function words, or because they are high frequency words, or both.

Several of the analyses performed on the database as a whole were performed again excluding the function words. For the 405 content words in the database, word frequency was not significantly correlated with error susceptibility ($r = -.0291$, $p = .560$). Familiarity and neighborhood characteristics of the words - density, mean neighborhood frequency, and prominence - were also not significantly correlated with error susceptibility for the 323 content words in the lexicon. For the 50 function words, however, word frequency was significantly negatively correlated with error susceptibility ($r = -.5452$, $p = .0001$) as was prominence ($r = -.5373$, $p = .0001$): as word frequency increased and as the frequency of the word relative to its neighbors increased, errors on function words decreased. Thus, content words and function words behave quite differently in terms of their susceptibility to error and in their lexical characteristics. Function words may have an advantage in intelligibility over content words because they are more predictable and because of redundant information in sentences, but they are generally shorter in duration and often acoustically reduced.

Sentence-level Characteristics

Total errors for each of the 100 sentences in the database were examined to see if characteristics of the sentence as a whole have any bearing upon word intelligibility. The monoclausal Harvard sentences are fairly homogeneous in syntactic structure, so the influence of sentence complexity upon intelligibility could not readily be assessed in this study.

Not surprisingly, there was a modest correlation between the number of words in a sentence and the number of errors for that sentence ($r = +.27$, $p < .008$). Sentences with more words tended to have more errors. There was no significant correlation, however, between the number of words in a sentence and the number of errors per word in that sentence ($r = .04$, $p < .709$). Therefore, increasing the length of the sentence does not necessarily result in an increase in the number of errors on words appearing within that sentence.

Twenty sentences were examined in greater detail: the 10 sentences with the highest number of errors and the 10 sentences with the lowest number of errors. The high-error sentences had an average of 121.2 errors per sentence, and the low-error sentences had an average of only 10.5 errors per sentence. The high-error sentences, with a mean of 8.3 words per sentence were significantly longer than the low-error sentences, with a mean of 6.7 words ($t = -4.028$, $df = 18$, $p <.001$). However, the dramatic increase in errors for the high error sentences cannot be accounted for solely by the addition of words. Examination of the words in these sentences revealed that high-error sentences were much more likely to contain high-error words than low-error sentences. The high-error sentences contained an average of 0.8 of the 10 words with the highest total error counts, whereas the low-error sentences contained an average of only 0.1 of the 10 highest-error words.

The transcription errors were also analyzed to determine if the position of the word within the sentence was related to the number of transcription errors occurring on that word. The results of a chi-square test comparing the expected number of total errors for each position in the sentence to the number of errors actually observed, revealed that words that occurred in the middle portion of sentences had more errors than expected whereas the initial two words of sentences had fewer errors than expected ($P^2 = 35.33$, df = 2, $p < .001$). The initial words appeared resistant to typing/spelling errors, omissions, semantic errors, and additions but were susceptible to consonant and vowel errors. The final two words of sentences, however, tended to be susceptible to typing/spelling errors and omissions while being relatively resistant to any type of phonetic error. Middle words may be susceptible to errors because they receive less attention or are not remembered as well as initial and final words. Initial words are probably resistant to most types of error because they are more prominent in terms of memory and attention. They may be susceptible to phonetic errors because the contextual cues present in the remaining portion of the sentence are unavailable for decoding the onset of the sentence. The susceptibility of final words to typing/spelling errors and omissions may represent failure of attention.

Talker Characteristics

Data for the 10 male and 10 female talkers were examined to determine if all talkers had similar amounts and types of errors. This analysis of talker-related error patterns allowed us to determine whether the overall patterns of listener errors are in any way related to talker-specific characteristics. As shown in Table II, the total number of errors for each talker ranged from 198 for Talker 19 to 488 for Talker 20 (mean number of errors = 305.5, SD = 84.44). To ascertain if patterns of errors were similar across speakers, chi-square tests comparing each talker's distribution of errors across the seven categories with the mean distribution of errors were performed. Fifteen of the 20 talkers had errors distributions which deviated significantly from the mean distribution. Talkers are not similar in number of transcription errors, and many talkers do not conform to the mean pattern of error distribution across error categories. Measures of central tendency, therefore, may not adequately characterize intelligibility for the range of normal talkers.

An ANOVA was performed to determine if female and male talkers differed in number of errors produced. Bradlow et al. (1996) found that the 10 female talkers in this database had significantly higher intelligibility scores than the 10 male talkers. In that study, sentences were counted as correctly transcribed if, and only if, all five "keywords" in the sentence were correct. Using the scoring method of counting errors for all words in the sentences, however, no significant difference in total errors between males and females was found ($F(1, 18) = 1.90$, $p < .185$). There was also no significant interaction between type of error and gender indicating that no specific error types were more common among one gender than the other ($F(7, 126) = 1.66$, $p < .124$).

To investigate whether some types of errors are more prevalent for talkers with higher total error scores, the percentage of total errors comprised by each error category was correlated with the total error score for each of the 20 talkers. As total error score increased, the percentage of typing/spelling errors increased significantly ($r = .6561$, $p < .002$), and the proportion of omitted words decreased significantly ($r = -.6488$, $p < .002$). The analysis was performed again excluding the typing/spelling errors: the only significant finding was a decrease in semantic errors as total errors increased. For talkers with greater total errors, no single category except typing/spelling errors is disproportionately responsible for the increase in total error.

Table 2.

Number of errors in each category for each of the 20 talkers.

Talker	Gender	Typing	Phonetic	Omitted	Semantic	Phon/Other	Addition	Unknown	Total
1	M	221	69	63	11	45	19	6	434
2	F	116	62	79	35	26	9	11	338
3.	M	75	52	51	20	28	8	9	243
4	F	61	46	60	22	18	8	7	222
5	F	71	60	72	25	17	7	6	258
6	M	90	96	47	31	44	10	77	325
7	M	140	93	49	31	44	9	7	373
8	M	52	91	62	29	17	6	4	261
9	M	70	98	38	20	23	4	5	258
10	M	67	59	73	14	14	7	10	244
11	F	127	145	63	50	35	16	9	445
12	F	58	50	71	18	28	6	3	234
13	F	160	51	92	27	25	4	14	373
14	F	66	44	65	5	15	8	7	210
15	M	83	103	59	35	18	10	7	315
16	F	69	92	44	27	33	7	8	280
17	M	106	82	98	25	36	11	13	371
18	F	57	61	72	13	27	9	6	245
19	F	64	51	39	18	18	7	3	198
20	M	24	126	40	46	37	7	8	488

The error patterns for the three least intelligible talkers were also examined more closely. Talker 1, who had a total of 434 errors, and Talker 20, with 488 total errors, both had high proportions of typing/spelling errors and added words. Talker 11, however, had a high percentage of phonetic and semantic errors among his 445 total errors. Thus, talkers with similar amounts of errors did not have similar proportions of errors spread across the error categories. These findings suggest that talkers who have reduced speech intelligibility may have reduced intelligibility for different reasons. The analysis of data by talker also suggests that general patterns of listener errors are related to segment and word characteristics independent of the talker.

Conclusions

Although most of the words and sentences in this database produced by normal talkers were transcribed accurately, it is clear from the 6100 transcription errors that the intelligibility of normal speech is not perfect. It is also clear that many factors affect normal speech intelligibility: speech intelligibility differs across talkers, segments and words. An earlier companion study by Bradlow et al. (1996) examined the talker-related acoustic-phonetic contributions to speech intelligibility. The present study used listener transcription errors to examine the contributions of the phonetic content and the lexical characteristics of the words to variability in intelligibility of normal speech.

The analysis of listeners errors revealed several systematic patterns that are related to the phonetic and lexical qualities of the words independent from talker characteristics. Regarding phonetic factors, we found that consonants were more susceptible to error than vowels. Further examination of consonant errors revealed that nonsonorant consonants were more likely to be transcribed incorrectly than sonorant consonants. In addition, consonant phonemes in the final position of words were more vulnerable to error, whereas those in the initial or medial positions of words were resistant to error. Consonants in clusters were more susceptible to consonant errors than singleton consonants that had fewer errors than expected.

The lexical status of words as content or function words was also an important factor in intelligibility. Content words were more susceptible to transcription error than function words. Several differences in the lexical characteristics of content and function words were noted that may have contributed to the difference in their observed error rates. Function words were significantly higher in frequency and were more prominent in their lexical neighborhoods than content words. The lower error rates for function words may result from their predictability in sentence context and because there are fewer words in the lexicon with which to confuse them.

Other lexical factors also exerted some influence upon the likelihood of transcription error for words. Verbs were more error-prone than other words, possibly because their word-final morphological endings are especially susceptible to misperception. Word length was not significantly correlated with error susceptibility. However, the shortest words in the database (one to two phonemes in length) were less likely to be incorrectly transcribed than was expected. These short words tend to be the more error-resistant function words rather than content words. Also, words with higher frequency and greater prominence in their lexical neighborhoods were less vulnerable to error than was anticipated based upon their distribution. These effects, however, are confounded with the content-function distinction between words.

There were no clear patterns of errors across the 20 talkers. This finding suggests that whereas the *number* of errors produced by a given talker (i.e., the talker's overall intelligibility score) may be related to acoustic-phonetic characteristics of his or her speech (Bradlow et al., 1996), the specific *types* of errors produced are more closely related to the phonetic and lexical characteristics of the words themselves.

Detailed analysis of highly error-prone words revealed that several characteristics may be responsible for increased susceptibility to transcription error: a single general-purpose explanation for word intelligibility does not exist. Some words were simply more difficult for listeners to type or spell. Semantic errors often occurred when one adjective was substituted for another. A few words provoked fairly consistent vowel substitutions, perhaps caused by differences in dialect between talkers and listeners. Combinations of words in sentences also led to problems in transcription: several phoneme sequences across word boundaries were difficult for listeners to accurately parse.

Taken together, the results of this extensive analysis of listener transcription errors provide insight into the complex relations among various phonetic and lexical factors that affect normal speech intelligibility. The patterns that emerged from our analysis suggest that all segments and words are not equal in intelligibility. It is interesting that consonant segments in word-final position and non-sonorant consonants are the most vulnerable to error for normal-hearing listeners in light of the findings of Picheny et al. (1986) regarding the acoustic differences between clear and conversational speech for hearing-impaired listeners. They found that clear speech is marked by more pauses between words, release of stops and other word-final consonants, and increased RMS intensities for obstruents, particularly stops. It appears that talkers employ their implicit knowledge of difficult-to-transmit segments in speaking more clearly for impaired listeners. In other words, the acoustic-phonetic comparisons between clear and conversational speech and the present analysis of normal listener transcription errors provide converging evidence regarding the speech sounds most vulnerable to transmission error. When talkers are asked to modify the way they speak, they selectively change precisely those aspects of the speech signal that are known to be prone to increased perceptual errors by listeners. These changes in articulation increase the likelihood that the listeners will recognize the intended message and recover the correct sequence of words. The present findings have several important theoretical implications. First, it is necessary to study speech intelligibility using speech samples from a number of talkers producing a large sample of speech. Second, it is necessary to study speech samples produced under different speaking conditions. Words produced in isolation are very different acoustically from words produced in fluent sentences. Finally, it is necessary to begin the difficult task of studying the many different sources of variability that affect speech perception and production. Much of what we currently know about speech communication has come from highly controlled laboratory experiments using a small number of cooperative talkers who produce speech samples using citation-form speech. Our knowledge of how speech is transformed and modified in other speaking environments and how these factors affect speech intelligibility remains a major challenge for the next few years.

References

Black, J.W. (1952). Accompaniments of word intelligibility. *Journal of Speech & Hearing Disorders*, 17, 409-418.

Black, J.W. (1957). Multiple-choice intelligibility tests. *Journal of Speech & Hearing Disorders*, 22, 213-235.

Bond, Z.S. & Garnes, S. (1980). *Misperceptions of fluent speech*. In R. A. Cole (Ed.), Perception and Production of Fluent Speech, Erlbaum: Hillsdale, NJ, pp. 115-132.

Bond, Z.S. & Moore, T.J. (1994). A note on the acoustic-phonetic characteristics of inadvertently clear speech. *Speech Communication*, 14, 325-337.

Byrd, D. (1994). Relations of sex and dialect to reduction. *Speech Communication*, 15, 39-54.

Hood, J.D. & Poole, J.P. (1980). Influence of the speaker and other factors affecting speech intelligibility. *Audiology*, 19, 434-455.

House, A.S., Williams, C.W., Hecker, M.H.L. & Kryter, K.D. (1965). Articulation testing methods: Consonantal differentiation with a closed response set. *Journal of Speech & Hearing Research*, 37, 158-166.

IEEE (1969). IEEE recommended practice for speech quality measurements. *IEEE Report No. 297.*

Karl, J. & Pisoni, D.B. (1994). The role of talker-specific information in memory for spoken sentences. *Journal of the Acoustical Society of America*, 95, 2873.

Kucera, F. & Francis, W. (1967). *Computational Analysis of Present Day American English*, Providence, RI: Brown University Press.

Luce, P.A. (1986). Neighborhoods of words in the mental lexicon. *Research on Speech Perception Technical Report No. 6*. Indiana University Speech Research Laboratory, Bloomington, In.

Luce, P.A. & Carrell, T.D. (1981). Creating and editing waveforms using WAVES. *Research on Speech Perception Progress Report No. 7* Indiana University Speech Research Laboratory, Bloomington, In.

Miller, G.A. & Nicely, P.E. (1955). An analysis of perceptual confusions among some English consonants. *Journal of Speech & Hearing Research*, 27, 338-352.

Nusbaum, H.C., Pisoni, D.B. & Davis, C.K. (1984). Sizing up the Hoosier mental lexicon: Measuring the familiarity of 20,000 words. *Research on Speech Perception Progress Report No. 10*, Indiana University Speech Research Laboratory, Bloomington, In.

Picheny, M.A., Durlach, N.I. & Braida, L.D. (1985). Speaking clearly for the hard of hearing I: Intelligibility differences between clear and conversational speech. *Journal of Speech & Hearing Research*, 28, 96-103.

Picheny, M.A., Durlach, N.I. & Braida, L.D. (1986). Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech. *Journal of Speech & Hearing Research*, 29, 434-446.

Picheny, M.A., Durlach, N.I. & Braida, L.D.(1989). Speaking clearly for the hard of hearing III: An attempt to determine the contribution of speaking rate to difference in intelligibility between clear and conversational speech. *Journal of Speech & Hearing Research*, 32, 600-603.

Pickett, J.M. & Pollack, I. (1963). Intelligibility of excerpts from fluent speech: effects of rate of utterance and duration of excerpt. *Language & Speech*, 6, 151-164.

Pisoni, D.B, Nusbaum, H.C., Luce, P.A. & Slowiaczek, L.M. (1985). Speech perception, word recognition, and the structure of the lexicon. *Speech Communication*, 4, 75-95.

Schultz, M.C. (1964). Word familiarity influences in speech discrimination. *Journal of Speech & Hearing Research*, 7, 395-400.

Some Observations on Neighborhood Statistics of Spoken English Words[1]

Shigeaki Amano[2]

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405

---

[1] This work is supported in part by NIH-NIDCD Research Grant DC00111 to Indiana University. I am grateful to Prof. David Pisoni and Gina Torretta who allowed me to use their database.

[2] NTT Basic Research Laboratories, 3-1 Morinosato Wakamiya, Atsugi, Kanagawa, 243-01 Japan.

475

Some Observations on Neighborhood Statistics of Spoken English Words

Abstract. Lexical neighborhoods were analyzed as a function of word length and phoneme position in terms of the neighborhood density, the mean neighborhood frequency, and the maximum neighborhood frequency. It was found that the neighborhood is very sparse in words more than six phoneme long, and that there are phoneme positional differences in the neighborhood, suggesting that the conventional lexical neighborhood may not be a good word candidate set for spoken word recognition.

Introduction

Research on spoken word recognition suggests that a word is processed interactively with other word candidates in a mental lexicon (e.g., McQueen, Norris, & Cutler 1994; Norris, McQueen, & Cutler 1995; Zwitserlood, 1989). Some recent models of spoken word recognition explicitly take account of the competition among word candidates. For example, the TRACE model (McClelland & Elman, 1986) and the SHORTLIST model (Norris, 1994; Norris, McQueen, & Cutler, 1995) incorporate lexical competition by adopting inhibitory connections among word candidates. On the other hand, the original cohort model (Marslen-Wilson & Welsh, 1978), the new cohort model (Marslen-Wilson, 1987), and the Neighborhood Activation Model (Luce, 1986) implicitly take account of the interactions among word candidates.

One of the findings used to support an interactive view is the effect of lexical neighborhood on word recognition. A neighborhood is a collection of words which have single phoneme substitution with a target word (e.g., Frauenfelder, 1990; Frauenfelder, Baayen, Hellwig, & Schreuder, 1993; Pisoni, Nusbaum, Luce & Slowiaczek, 1985).

Precisely speaking, there is another definition for the neighborhood. This definition includes words with single phoneme deletion or addition to the target words in addition to the substitution (e.g., Goldinger, 1989; Luce, 1986; Sommers, 1996). However, the former is used as the definition of neighborhood in this paper, because it is simpler than the latter and no substantial differences were found in the statistical characteristics of the two definitions (Frauenfelder, Baayen, Hellwig, & Schreuder, 1993).

The characteristics of the neighborhood have been described by three variables in previous studies. The first variable is the "density" which is the number of words in the neighborhood (e.g., Frauenfelder, Baayen, Hellwig, & Schreuder, 1993; Luce, 1986). The second variable is the "mean frequency" which is the averaged frequency of words in the neighborhood (e.g., Frauenfelder, Baayen, Hellwig, & Schreuder, 1993; Luce, 1986). The third variable is the "maximum frequency" which is the highest frequency of a word in the neighborhood. (Bard, 1990; Bard & Shillcock, 1993).

Several studies have shown that the neighborhoods have significant effects on word recognition. For example, neighborhood density and frequency negatively correlate with recognition rate of a target word (Luce, 1986). Neighborhood density has inhibitory effects on reaction times of lexical decision and naming for a target word (Goldinger, 1989). Low frequency words in neighborhood have negative priming effects on target word recognition (Goldinger, Luce, & Pisoni, 1989). Recognition of two-syllable target words (spondees) is affected by neighborhood characteristics of each syllable in the spondee (Cluff & Luce, 1990). And, older adults have difficulty recognizing a target word if density and frequency is high in neighborhood (Sommers, 1996).

One of the problems in these studies is that almost all of them used only monosyllabic words. The exception was Cluff and Luce (1990) who used two-syllable spondee words. However, they calculated the neighborhood for each syllable not for a entire word. Therefore, effects of neighborhood have not been investigated in multisyllabic words. The reason is probably that the neighborhood is not effective in the multisyllabic words. In the initial analysis in this study, this idea will be confirmed by examining the statistics of neighborhoods.

Another problem with the previous studies is concerned with an assumption concerning neighborhood characteristics. In the past, researchers have assumed that neighborhoods at each phoneme position within a word have equivalent effects on word recognition. That is, the conventional definition of a neighborhood does not take into account positional factors. Frauenfelder, Baayen, Hellwig, and Schreuder (1993) have already mentioned this point, but did not conduct any analyses to investigate these effects. If speech is processed in a left-to-right manner (Cole & Jakimik, 1980) and if a spoken word is recognized by reference to the mental lexicon in a left-to-right manner as some word recognition models claim (e.g., Marslen-Wilson & Welsh, 1978; McClelland & Elman, 1986), then each phoneme position may contribute to word recognition differently. For instance, the first position might have a larger contribution to word recognition than other phoneme positions in a word.

Amano, Torretta, and Luce (1997) have found positional neighborhood effects. They conducted correlation analyses between word identification rate and three neighborhood variables; the density, the mean log frequency, and the maximum log frequency at each phoneme position in CVC English words. They reported that the first phoneme position had greater negative effects on word recognition than any of the other phoneme positions.

However, Amano, Torretta, and Luce (1997) used CVC words in their study. As a consequence, the length of words was restricted to only three phonemes. It is unknown whether their results are applicable to other word lengths. A second analysis focused on this point in order to provide indirect evidence that the first phoneme position is also effective in other word lengths. Thus, the statistics of the positional neighborhood was examined in longer words.

Analysis 1

Method

A computerized dictionary (Nusbaum, Pisoni, & Davis, 1984) was used for all the following analyses. The dictionary contains 19,295 words with Kucera and Francis (1967) word count. However, only 19,152 words with two to twelve phoneme long were used for the following analyses, because there is no neighborhood for words with more than 13 phoneme long. The analyses were conducted as a function of word length. Independent variables were the number of words, the log frequency of a target word, the neighborhood density, the mean neighborhood log frequency, and the maximum neighborhood log frequency.

Results and Discussion

Figure 1 shows the distribution of the number of words as a function of word length. It shows that the words which are five phoneme long are the most frequent. The two phoneme long words and words in more than 10 phoneme long are not very frequent.

Insert Figure 1 about here.

Figure 2 shows the distribution of log frequencies of a target word. The results indicate that two phoneme words are the most frequent and that word frequency exponentially decreases as word length increases.

Insert Figure 2 about here.

Figure 3 shows the distribution of neighborhood density as a function of word length. The neighborhood is very dense for two and three phoneme words, but becomes very sparse in words with more than six phonemes. Note that monosyllabic words are up to seven phonemes long in English and that such words are frequent in three and four phoneme long. These results suggest that neighborhoods are present for monosyllabic words but are virtually non-existent for multisyllabic words. Therefore, the neighborhood is only effective for monosyllabic words but not for multisyllabic words.

Insert Figure 3 about here.

This conclusion is supported by the distribution of the mean neighborhood log frequency (Figure 4) and the maximum neighborhood log frequency (Figure 5). These two variables are almost zero for words with more than six phonemes. The findings suggest that neighborhood frequency is not effective for those words. Therefore, neighborhood is not effective either in terms of density or frequency for multisyllabic words.

However, as shown by Cluff and Luce (1990), if neighborhood is calculated at each syllable in multisyllabic words, it affects recognition of multisyllabic words. Therefore, the effects of neighborhood are not straightforward for multisyllabic words. It is effective via a syllable but not via an entire word.

Insert Figure 4 about here.

Insert Figure 5 about here.

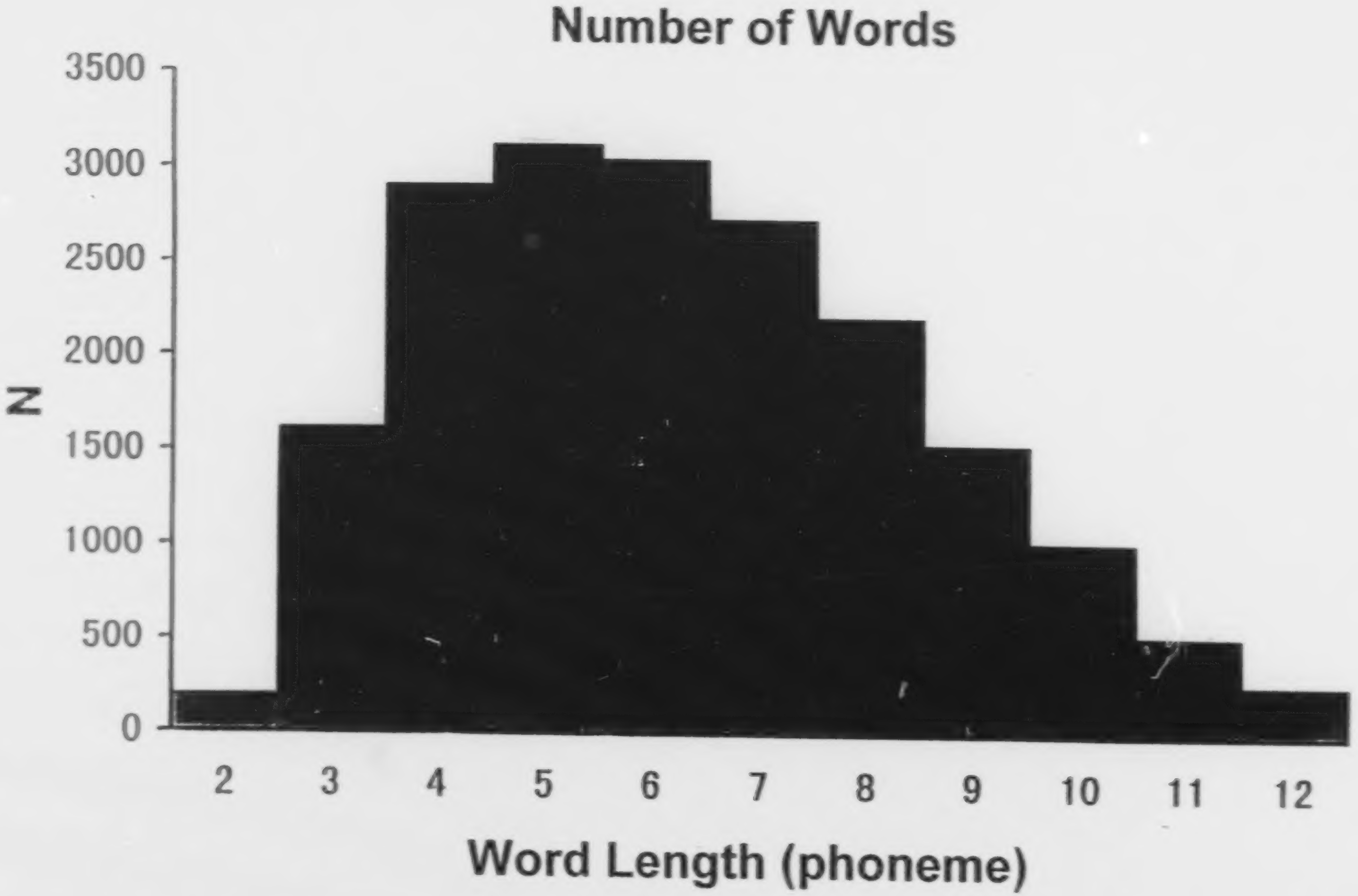


Figure 1. The number of words as a function of word length.

BEST COPY AVAILABLE

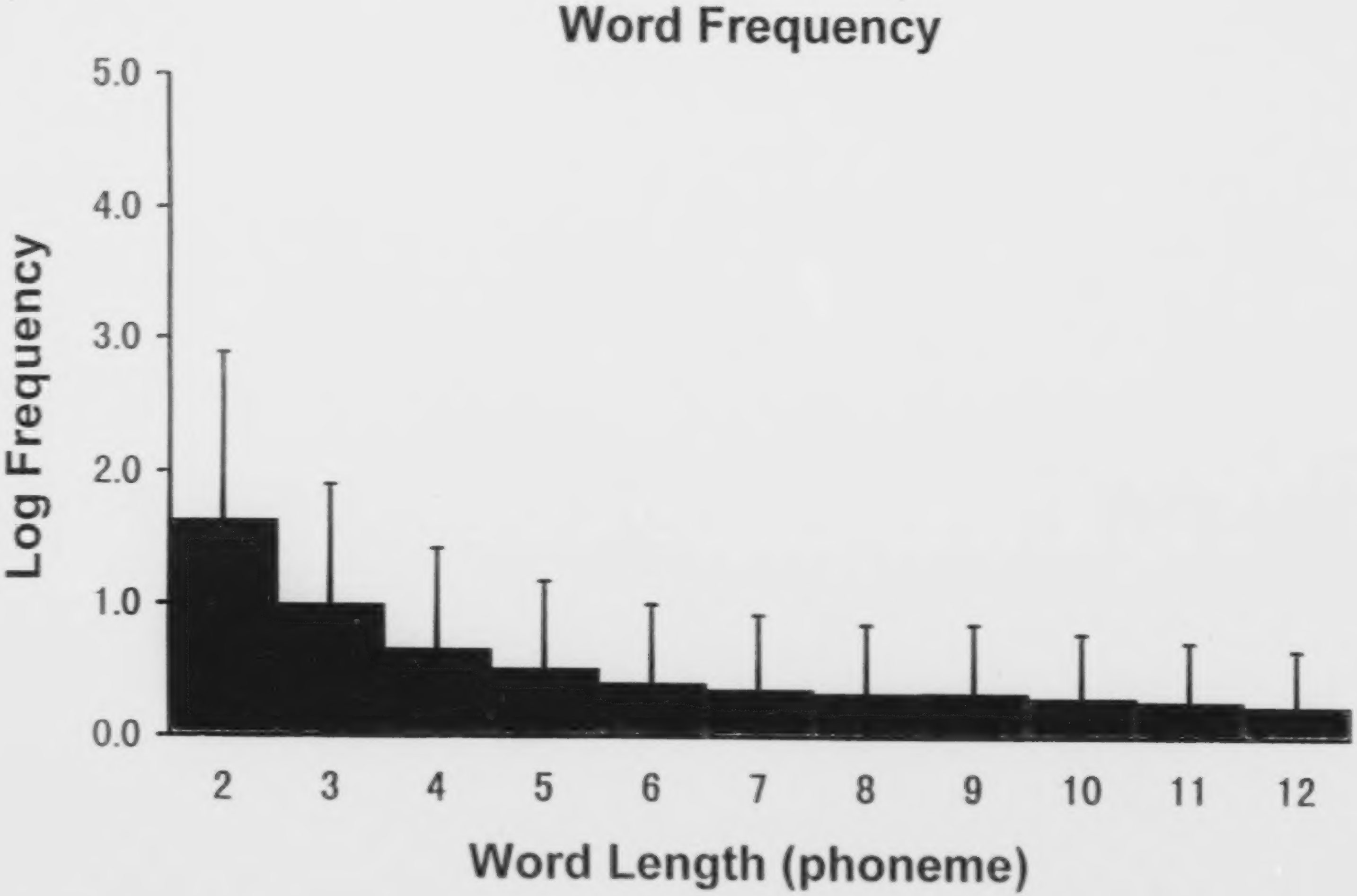


Figure 2. Word frequency of a target word as a function of word length.

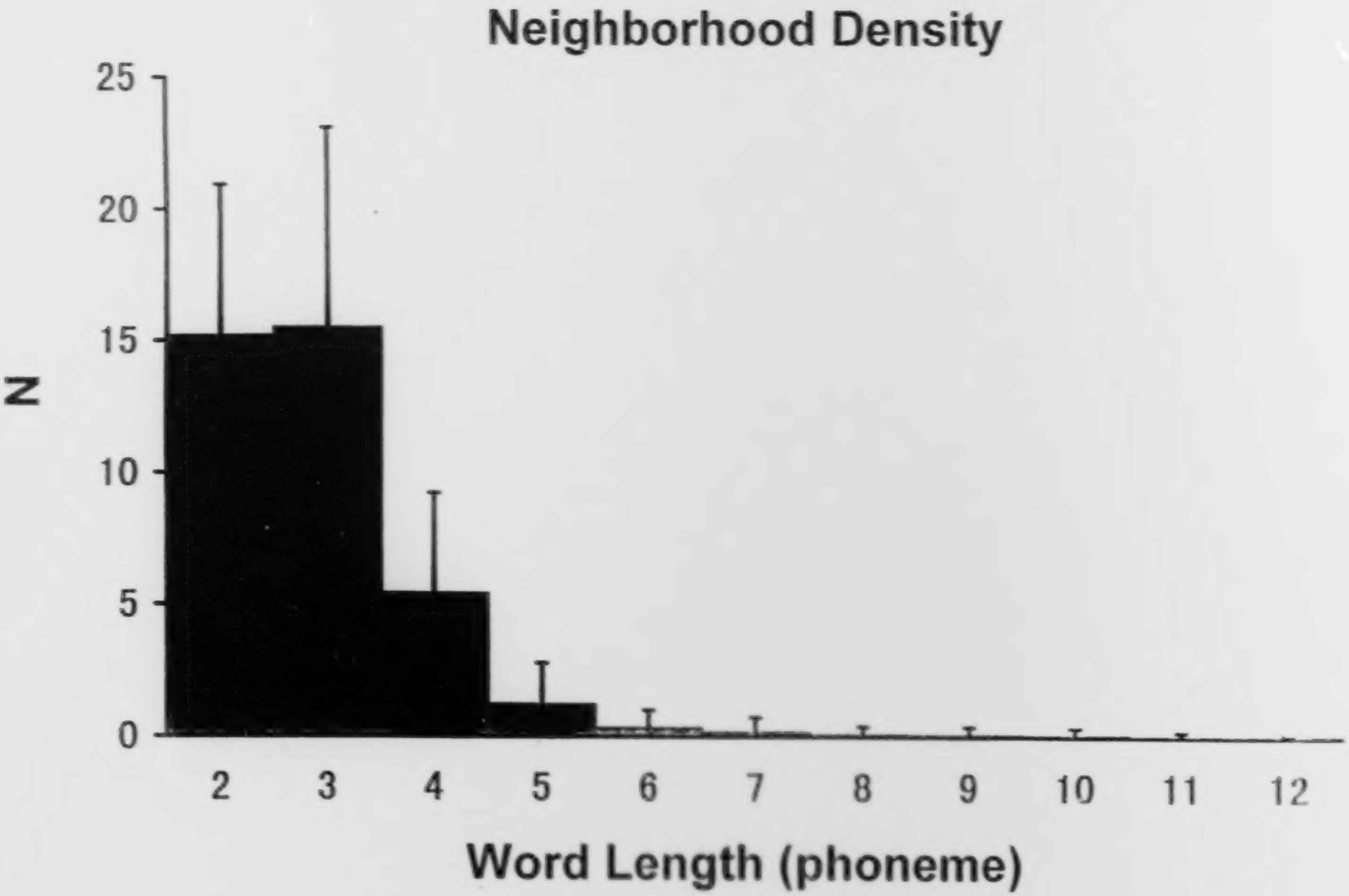


Figure 3. Neighborhood density as a function of word length.

BEST COPY AVAILABLE

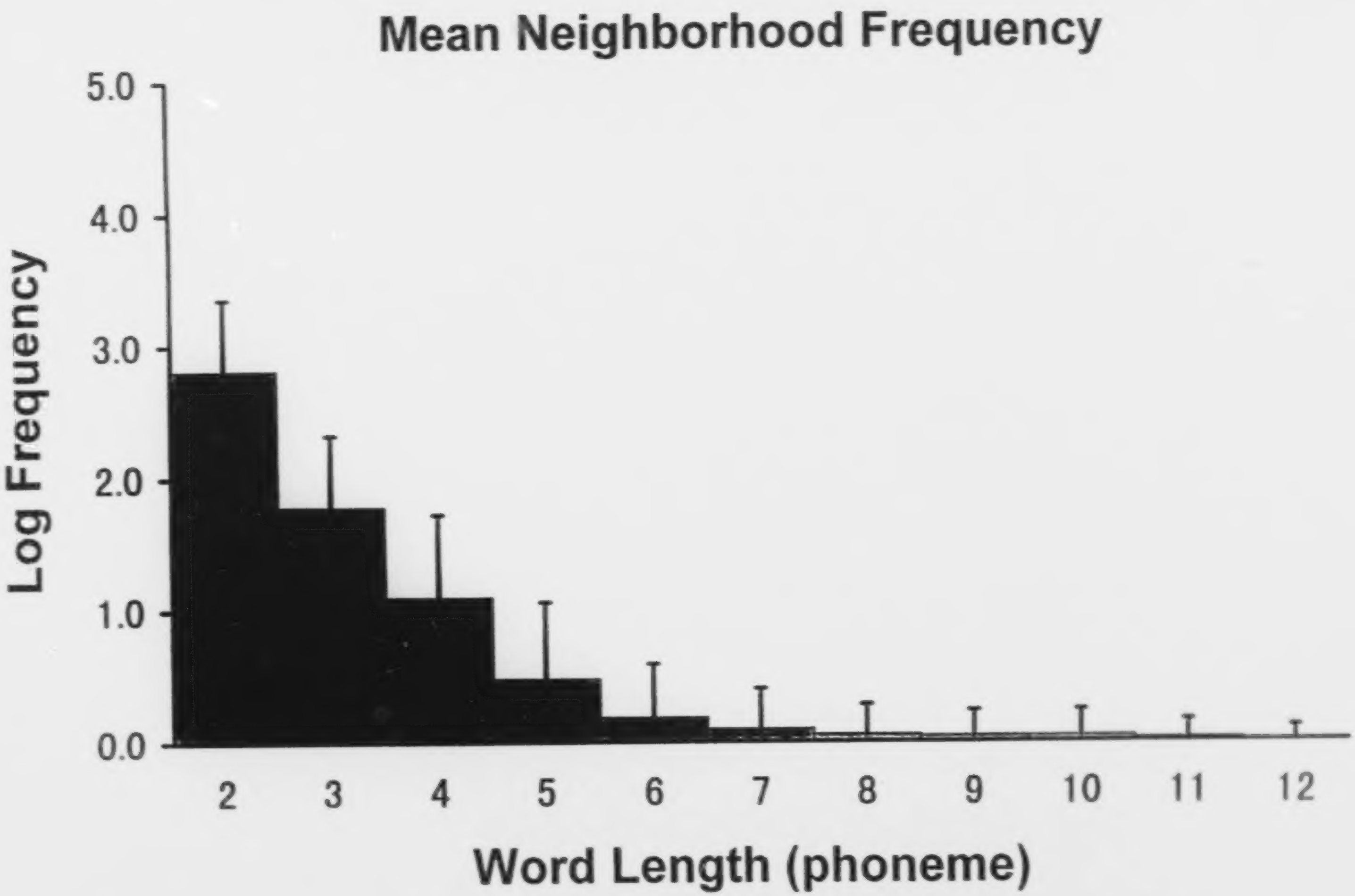


Figure 4. Mean neighborhood log frequency as a function of word length.

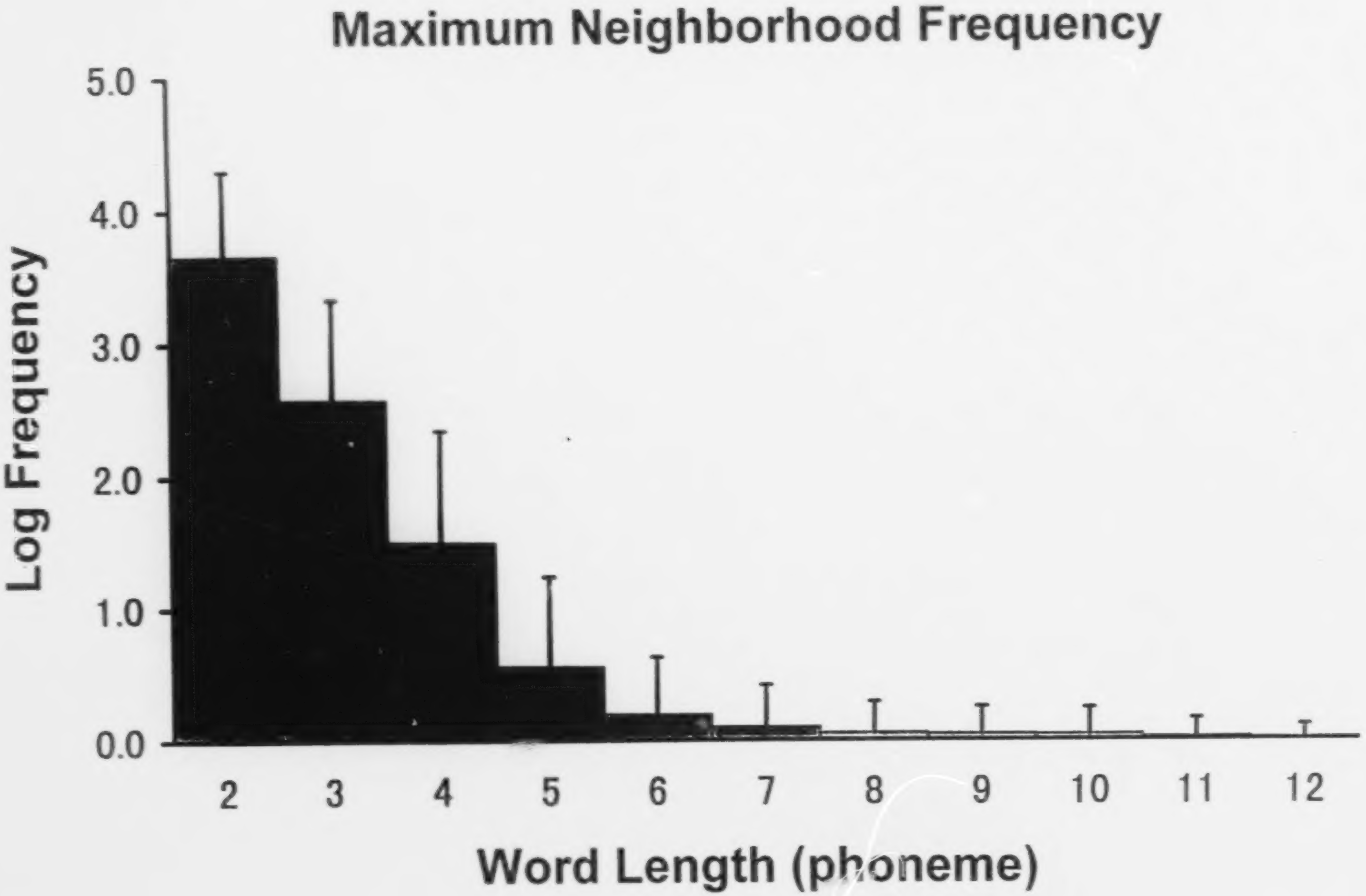


Figure 5. Maximum neighborhood log frequency as a function of word length.

Analysis 2

Method

The same word set as Analysis 1 was used for another series of analyses. This analysis was concerned with statistics of positional neighborhoods (Amano, Torretta, & Luce, 1997) which have single phoneme substitution with a target word at a particular phoneme position. For example, if a target word is "cat," then "bat," "fat," "hat," and "pat" are in the positional neighborhood at the first phoneme position. The density, the mean log frequency, and the maximum log frequency of the positional neighborhood were calculated for every phoneme position in two to twelve phoneme long words.

Results and Discussion

Figures 6, 7, and 8 show the distribution of the density, the mean log frequency, and the maximum log frequency of the positional neighborhood. Like the conventional neighborhood which was examined in Analysis 1, the density of the positional neighborhood is very low for words with more than six phonemes (Figure 6). The mean log frequency (Figure 7) and the maximum log frequency (Figure 8) of positional neighborhood are also very low for these words. Therefore, it is suggested that the positional neighborhood is not effective for words with more than six phonemes. This finding suggests that the positional neighborhood is not as effective in multisyllabic words as the conventional neighborhood as shown in Analysis 1.

Amano, Torretta, and Luce (1997) have shown that the positional neighborhood at the first phoneme position is more effective as a predictor for lexical competition than that at other phoneme positions. They used three phoneme words with CVC pattern.

The present analyses showed that there was a significant difference among the phoneme positions in three phoneme long words in terms of the neighborhood density, $F(2,4896) = 242.69$, $p < .0001$, the mean neighborhood log frequency, $F(2,4896) = 19.54$, $p < .0001$, and the maximum neighborhood log frequency, $F(2,4896) = 46.21$, $p < .0001$. The HSD test between a pair of positions showed that the first phoneme position is greater than the other phoneme positions in almost all cases in terms of all neighborhood variables with 5% significance level. The only one exception was that there was no significant difference between the first phoneme position and the third phoneme position in term of the mean neighborhood log frequency.

The results of Amano, Torretta, and Luce (1997)'s study might be due to the superiority of the first phoneme position in the positional neighborhood. In that case, the effect of positional neighborhood at the first phoneme position on spoken word recognition may be observed not only in three phoneme words but also in words of other lengths.

Insert Figure 6 about here.

Insert Figure 7 about here.

Insert Figure 8 about here.

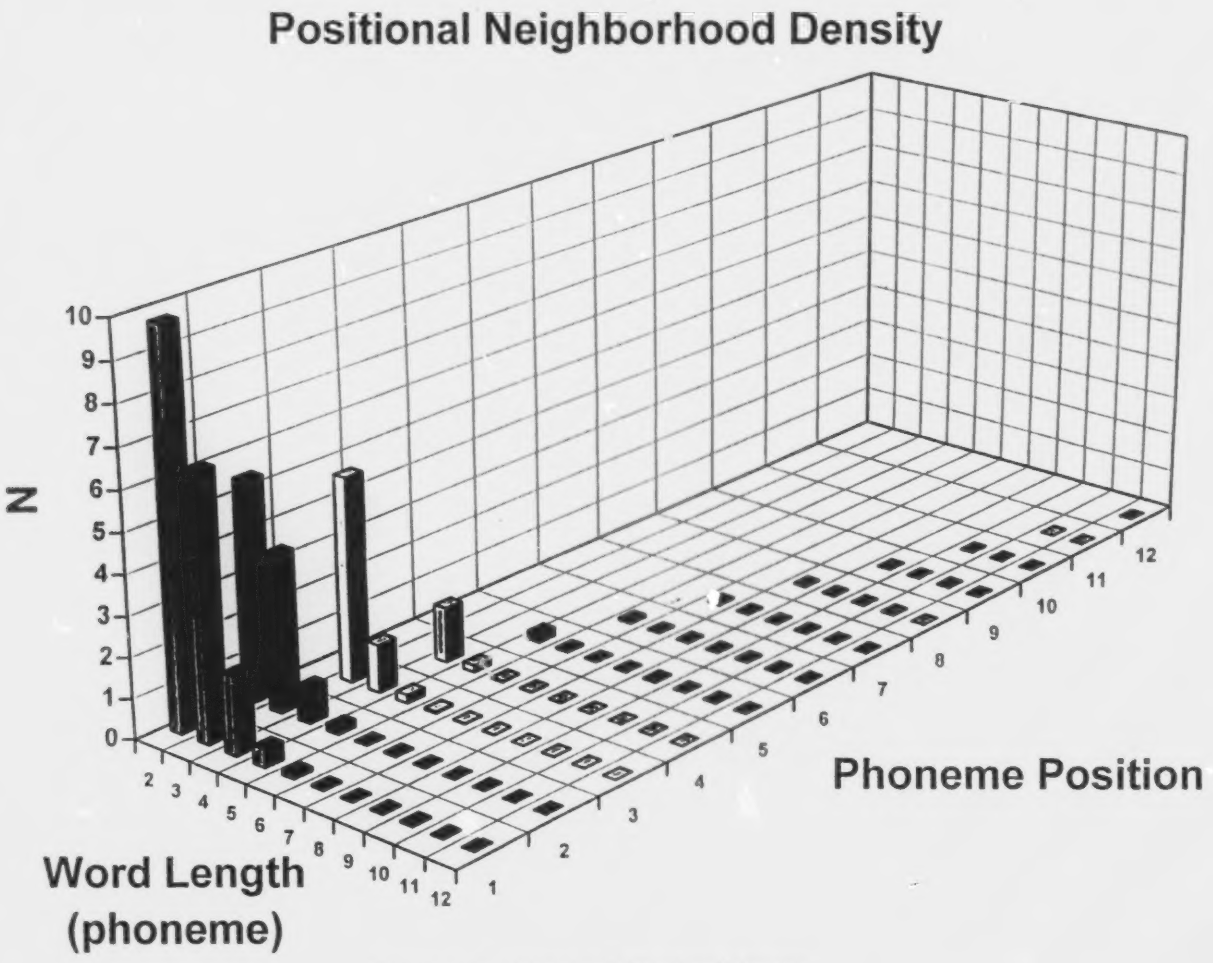


Figure 6. Positional neighborhood density.

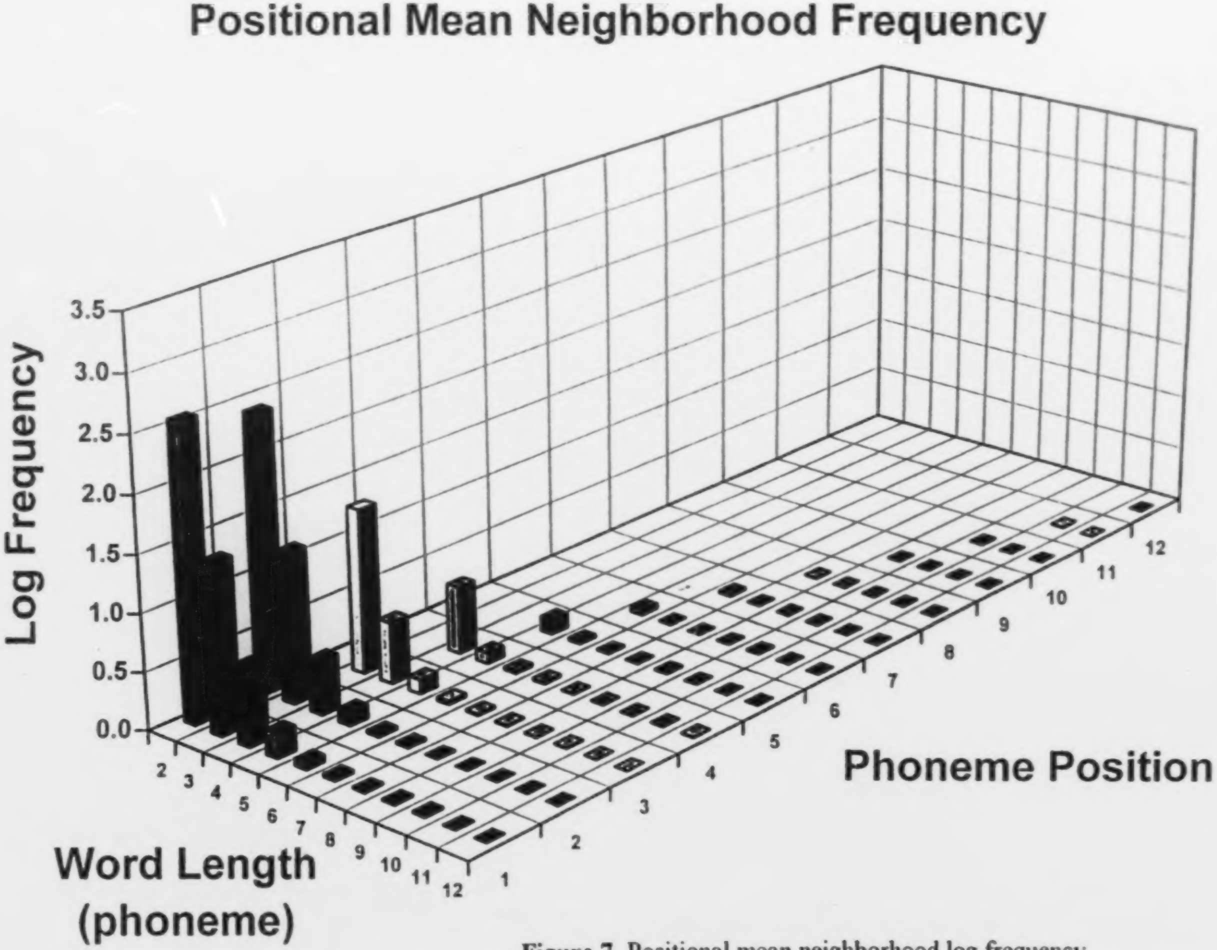


Figure 7. Positional mean neighborhood log frequency.

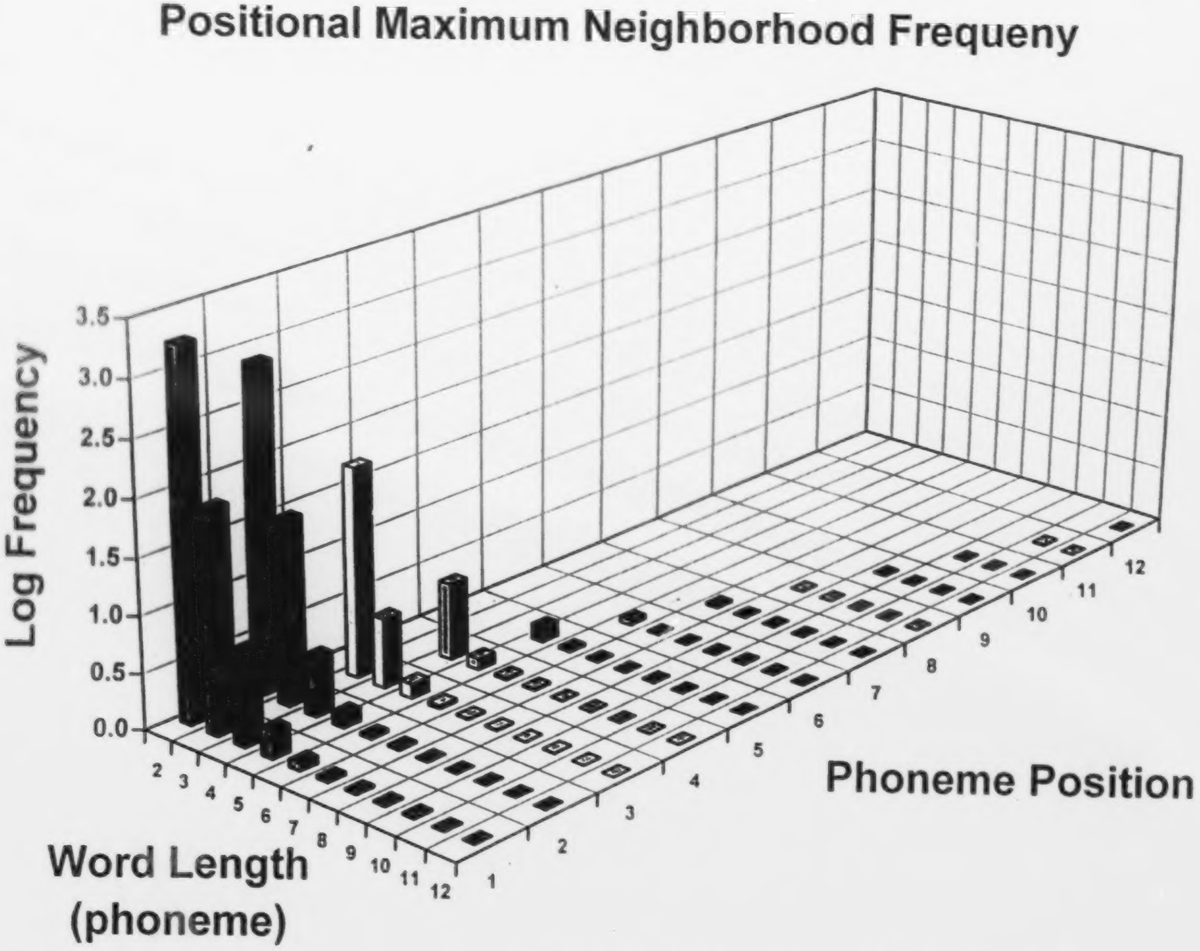


Figure 8. Positional maximum neighborhood log frequency.

BEST COPY AVAILABLE

Support for this proposal was obtained as follows. There was a significant main effect of phoneme position in almost all lengths with 1% significance level (exceptions were 12 phoneme words for the neighborhood density, 2, 11, and 12 phoneme words for the mean neighborhood log frequency, and 11, and 12 phoneme words for the maximum neighborhood log frequency). In addition, the first phoneme position is significantly greater than all other phoneme positions in two to seven phoneme words in terms of the neighborhood density, in three to six phoneme words in terms of the mean neighborhood log frequency, and in two to six phoneme words in terms of the maximum neighborhood log frequency. This means that the first position has greater value than other phoneme positions in the word length where the positional neighborhood is virtually effective. Therefore, it is suggested that the first phoneme position contributes more to spoken word recognition than other phoneme positions not only in three phoneme words (Amano, Torretta, & Luce, 1997) but also in words of other lengths.

Conclusion

Analysis 1 suggested that the neighborhood is not effective in words with more than six phonemes, because the neighborhood is very sparse in those words. Analysis 2 showed that there are phoneme positional differences in the neighborhood and that the first phoneme position has greater value than other phoneme positions in terms of the neighborhood density, the mean neighborhood frequency, and the maximum neighborhood frequency.

These findings indicate that the conventional neighborhood has some difficulties in explaining lexical competition in spoken English words. It cannot explain the competition between long words, and it does not take into account positional differences. Therefore, the neighborhood is not suitable for the word candidate set for spoken word recognition. Further research is required to find a better word candidate set.

References

Amano, S., Torretta, G. M., & Luce, P. A. (1997). Positional neighborhood effects on spoken word recognition. *The Journal of the Acoustical Society of America*, **101**, Part 2, 4aSC9, 3155.

Bard, E. G. (1990). Competition, lateral inhibition, and frequency: Comments on the chapters of Frauenfelder and Peeters, Marslen-Wilson, and others. In G. T. M. Altmann (Ed.), *Cognitive models of speech processing: Psycholinguistic and computational perspectives* (pp. 185-210). Cambridge: MIT Press.

Bard, E. G., & Shillcock, R. C. (1993). Competitor effects during lexical access: Chasing Zipf's tail. In G. T. M. Altmann, & R. Shillcock (Eds.), *Cognitive models of speech processing: The second Sperlonga meeting* (pp. 235-275). Hillsdale: LEA.

Cluff, M. S., & Luce, P. A. (1990). Similarity neighborhoods of spoken two-syllable words: Retroactive effects on multiple activation. *Journal of Experimental Psychology: Human Perception and Performance*, **16**, 551-563.

Cole, R. A., & Jakimik, J. (1980). A model of speech perception. In R. A. Cole (Ed.), *Perception and production of fluent speech* (pp. 133-164). Hillsdale: LEA.

Frauenfelder, U. H. (1990). Structure and computation in the mental lexicon. In H. Haken & M. Stadler (Eds.), *Synagetics of cognition* (pp. 406-414). Berlin: Springer-Verlag.

Frauenfelder, U. H., Baayen, R. H., Hellwig, F. M., & Schreuder, R. (1993). Neighborhood density and frequency across languages and modalities. *Journal of Memory and Language*, **32**, 781-804.

Goldinger, S. D. (1989). Neighborhood density effects for high frequency words: Evidence for activation-based models of word recognition *Research on Speech Perception, Progress Report 15*, pp. 163-186. Bloomington, IN: Indiana University, Speech Research Laboratory.

Goldinger, S. D., Luce, P. A., & Pisoni, D. B. (1989). Priming lexical neighbors of spoken words: Effects of competition and inhibition. *Journal of Memory and Language*, **28**, 501-518.

Kucera, H., & Francis, W.N. (1967). Computational analysis of present-day American English. Providence: Brown University Press.

Luce, P. A. (1986). Neighborhoods of words in the mental lexicon. *Research on Speech Perception, Technical Report* **6**. Bloomington IN: Indiana University, Speech Research Laboratory.

Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word-recognition. *Cognition*, **25**, 71-102.

Marslen-Wilson, W. D., & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology*, **10**, 29-63.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, **18**, 1-86.

McQueen, J. M., Norris, D., & Cutler, A. (1994). Competition in spoken word recognition: Spotting words in other words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **20**, 621-638.

Norris, D. (1994). Shortlist: a connectionist model of continuous speech recognition. *Cognition*, **52**, 189-234.

Norris, D., McQueen, J. M., & Cutler, A. (1995). Competition and segmentation in spoken word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **21**, 1209-1228.

Nusbaum, H. C., Pisoni, D. B., & Davis, C. K. (1984). Sizing up the Hoosier mental lexicon: Measuring the familiarity of 20,000 words. *Research on Speech Perception, Progress Report 10*, pp. 357-376. Bloomington, IN: Indiana University, Speech Research Laboratory.

Pisoni, D. B., Nusbaum, H. C., Luce, P. A., & Slowiaczek, L. M. (1985). Speech perception, word recognition and the structure of the lexicon. *Speech Communication*, **4**, 75-95.

Sommers, M. S. (1996). The structural organization of the mental lexicon and its contribution to age-related declines in spoken-word recognition. *Psychology and Aging*, **11**, 333-341.

Zwitserlood, P. (1989). The locus of the effects of sentential-semantic context in spoken-word processing. *Cognition*, **32**, 25-64.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 21 (1996-1997)
Indiana University

Sensory Aid and Word Position Effects on Consonant Feature Production by Children with Profound Hearing Impairment[1]

Steven B. Chin,[2] Karen Iler Kirk,[2] and Mario A. Svirsky[2]

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405

---

[1] This research was supported by NIH/NIDCD Research Grant DC00423 to the Indiana University School of Medicine. A version of this paper was presented at the Vth International Cochlear Implant Conference, May 1997, New York, NY. We are grateful to Erin Diefendorf, Nicole Jones, Theresa Kerr, and Ted Meyer, all at the DeVault Otologic Research Laboratory of the Indiana University School of Medicine, for their assistance with this project. We also acknowledge staff at the St. Joseph Institute for the Deaf (St. Louis, MO), the University of Michigan Medical Center (Ann Arbor, MI), Boys Town National Research Hospital (Omaha, NE), and the Louisville Deaf-Oral School (Louisville, KY) for allowing children at their institutions to participate in this study.

[2] Also DeVault Otologic Research Laboratory, Department of Otolaryngology–Head and Neck Surgery, Indiana University School of Medicine, Indianapolis, IN.

S.B. Chin, K.I. Kirk, and M.A. Svirsky

Sensory Aid and Word Position Effects on Consonant Feature Production by Children with Profound Hearing Impairment

Abstract. This study examined consonant feature production by pediatric Nucleus 22-channel cochlear implant users (mean PTA = 110 dB) at two intervals: (1) within a few months before or after initial stimulation (mean = 0.36 years after) and (2) after approximately two years of device use. Additionally, the consonant feature production of the cochlear implant users at the later interval was compared with production by tactile aid users (PTA > 110 dB) and two groups of hearing aid users with differing amounts of residual hearing: those with mean PTAs of 93 dB (PTA_{93} hearing aid users) and those with mean PTAs of 104 dB (PTA_{104} hearing aid users). All four groups were had approximately the same average age at testing. All children were administered the *Goldman-Fristoe Test of Articulation*, which was scored in terms of percent correct voicing, place of articulation, and manner of articulation for consonants produced in word-initial, word-medial, and word-final positions. A longitudinal within-group comparison for the cochlear implant users revealed significant improvements in consonant feature production after approximately two years of device use. Consonant feature production for the cochlear implant users at the later interval surpassed that of age-matched tactile aid users and PTA_{104} hearing aid users, although PTA_{93} hearing aid users still outperformed all other groups. Finally, all groups exhibited word-position effects on consonant feature production, so that correct place, manner, and voicing production was highest in word-initial position, reduced in medial position, and lowest in word-final position. The results thus demonstrate the relative usefulness of cochlear implants for promoting appropriate speech production in deaf children, as well as the importance of linguistic variables (e.g., word-position) in assessing speech production performance.

Introduction

The effects of profound deafness on the speech production of postlingually deafened adults are minimal (e.g., Leder & Spitzer, 1990), but the effects on prelingual children can be devastating. Not only is hearing itself affected adversely, but the child's ability to acquire and use a spoken language is severely diminished. Thus, although cochlear implants were first developed for use by postlingually deafened adults, many investigators believe that the patient group that stands to benefit the most from cochlear implants will be "children, particularly young children" (Berliner, Eisenberg, & House, 1985). Such benefits would include, most importantly, provision of the necessary auditory input for the acquisition of target-appropriate speech production and spoken language.

Compared with research on the efficacy of cochlear implants in aiding auditory perception generally and speech perception specifically, the study of speech production by children using cochlear implants is relatively recent. Tobey, Angelette, et al. (1991) examined imitative segmental and nonsegmental characteristics, phonological skills, and intelligibility (following procedures developed by Ling, 1976, and McGarr, 1983) in 61 children who used the Nucleus multichannel cochlear implant. Tobey, Angelette, et al. reported that 79% of the children studied improved on a least one third of the measures. Improvement was most common on tasks assessing imitative segmental characteristics (66.7% of the children), followed by tasks examining intelligibility (62.9%), phonological skills (55.6%), and nonsegmental characteristics (31.1%). Similarly, Tobey and Hasenstab (1991) examined speech production by 78 users of the Nucleus device once preoperatively and up to four times postoperatively. These children demonstrated increases in scores on both segmental and suprasegmental measures postoperatively and with increased device use. Postoperatively, speech intelligibility was higher, but mean length of utterance was not significantly different. Tobey, Pancamo, Staller, Brimacombe, and Beiter (1991) examined consonant production in 29 children before they were fitted with a Nucleus device and again after one year of device use. A greater number of children produced stops, nasals, fricatives, and glides after implantation than before. After implantation, voiced stops were used by more children than were voiceless stops, but voiceless fricatives were produced more than voiced ones. Additionally, consonants with visible places of articulation were used more than those with less visible places.

It should be noted that children without cochlear implants may also show some improvements on the variables just discussed. Ultimately, it would be important to assess the improvement that is due to the implant alone, and not to maturation. To this end, more recent studies have introduced the use of control data from other populations (e.g., children using other sensory devices). Tobey, Geers, and Brenner (1994) analyzed the speech production skills of three groups (users of cochlear implants, tactile aids, and hearing aids) of 13 children matched by age, hearing loss, intelligence, family support, and speech and language skills (Geers & Moog, 1994). All children had better ear PTA thresholds of 100 dB HL or greater[3] and were tested once a year for three years in both imitative and spontaneous speech tasks. In addition, thirteen children with PTAs between 90 and 100 dB HL were tested once at the end of the study for comparison with the other three groups. For imitated speech production, significant differences among groups appeared first at the 24-month interval, when performance by cochlear implant users was better than that of tactile aid and hearing aid users on suprasegmentals and vowels/diphthongs. By the 36-month interval, cochlear implant users showed significantly better performance on most measures. For spontaneous speech, the cochlear implant users also showed significantly greater improvement in the production of both consonants and vowels than did the users of tactile aids and hearing aids. After three years of device use, the cochlear implant group showed similar performance to the children with PTAs between 90 and 100 dB HL.

Kirk, Diefendorf, Riley, and Osberger (1995) compared consonant feature production in CV syllables by 24 multichannel cochlear implant users at two intervals and further compared this with production by two groups of hearing aid users (16 children whose mean unaided PTA was 103 dB HL and 16 children whose mean unaided PTA was 94 dB HL). Cochlear implant users demonstrated significant improvements in the production of voicing, place, and manner features after approximately 2.6 years of device use. Additionally, both cochlear implant users and hearing aid users with a mean PTA of 94 showed significantly better place and voicing scores than did hearing aid users with a mean PTA of 103.

Two recent studies have compared the production of cochlear implant users and vibrotactile aid users. Ertmer, Kirk, Sehgal, Riley, and Osberger (1997) examined longitudinal changes in imitative vowel and diphthong production in 10 children using cochlear implants and 10 children using tactile aids. Production was evaluated at two intervals: (1) before children received a cochlear implant or tactile aid and (2) after at least one year of using the sensory aid ($M = 1.8$ years). From the earlier interval to the later interval, cochlear implant users showed significant improvement on seven of nine vowel and diphthong production measures, whereas the tactile aid users significantly increased performance on only one measure. Additionally, at the postdevice interval, cochlear implant users had significantly higher scores than the tactile aid users on eight of the nine measures. Sehgal, Kirk, Svirsky, Ertmer, and Osberger (1998) examined consonant feature production in CV syllables by cochlear implant users and vibrotactile aid users. Both groups were tested before receiving their devices and again approximately 1.5 years after receiving their devices. Users of both cochlear implants and tactile aids showed relatively poor production of voicing, place, and manner features at the predevice interval. Both showed improved production postdevice, but the improvement demonstrated by the cochlear implant users was significantly greater than that of the vibrotactile aid users. Cochlear implant users improved performance on one place feature and all of the manner features.

---

[3] Cochlear implant users had an average unaided threshold of 118 dB HL; both tactile aid and hearing aid users had an average unaided threshold of 110 dB HL.

Previous research has provided strong evidence that cochlear implants benefit the acquisition of certain aspects of speech production and spoken language in young children with prelingual profound hearing impairments. However, previous studies have either (1) relied exclusively on imitative production tasks or (2) limited sampling to consonants in restricted phonological environments or (3) lacked comparison with other groups of children. The present study examined elicited consonant feature production in pediatric cochlear implant users at two intervals and compared production at the later interval with production by age-matched tactile aid users and two groups of conventional hearing aid users. Speech samples were collected using the *Goldman-Fristoe Test of Articulation* (GFTA; Goldman & Fristoe, 1972), a standard instrument used in assessing articulation disorders and articulation development. Using these speech materials, consonant feature production was compared longitudinally for the cochlear implant users (early vs. late interval), as well as cross-sectionally among the late-interval cochlear implant users and the users of tactile aids and conventional hearing aids. Finally, productions of consonant features located in various word positions were compared in the various groups of children.

Method

Subjects

Subjects for this study were nine pediatric users of the Nucleus-22 multichannel cochlear implant (Patrick & Clark, 1991). Age at onset of deafness ranged from 0.0 years to 1.3 years (*M* = 0.41 years, *SD* = 0.56). Age at fitting with the cochlear implant ranged from 3.1 to 8.5 years (*M* = 5.38 years, *SD* = 1.52), so that the length of auditory deprivation ranged from 1.8 to 7.2 years (*M* = 4.97 years, *SD* = 1.43).

Consonant feature production of these children with cochlear implants was examined (see below) at two intervals: (1) Early interval: either immediately prior to implantation or within several months after implantation (years of device use: *M* = 0.36 years, *SD* = 0.25 years), and (2) Late interval: between 1.5 years and 2.6 years after implantation (*M* = 2.1 years, *SD* = 0.5 years) (late interval). For the nine cochlear implant users, the early interval occurred before implantation for three of the children, and within 0.6 years after implantation for the remaining six. Of the six children whose early interval occurred after implantation, five used the MPEAK strategy implemented on the MSP (mini speech processor) and one used the F0/F1/F2 strategy implemented on the WSP (wearable speech processor). Also for these six, the number of active electrodes ranged from 9 to 20, with five children using 12 or more active electrodes. At the late interval, eight of the children used the MPEAK strategy and one the F0/F1/F2 strategy. For these nine children, the number of active electrodes ranged from 5 to 20, with seven children using 12 or more active electrodes. Two of the children with cochlear implants used Oral Communication, and the remaining seven used Total Communication.

The consonant feature production of cochlear implant users at the late interval was compared with that of children using either tactile aids or conventional hearing aids. The nine children who used a tactile aid (Tactaid 7; see Franklin, 1991) were prelingually profoundly deaf and had PTAs > 110 at the time of testing. The hearing aid users were divided into two groups according to their unaided thresholds at 500, 1000, and 2000 Hz (Miyamoto, Osberger, Todd, & Robbins, 1994; Osberger, Maso, & Sam, 1993). Nine children who used hearing aids demonstrated hearing levels of 90-100 dB HL at two of the three frequencies with none of the thresholds higher than 105 dB HL; mean PTA for this group was 93 dB HL. These children were designated as the PTA_{93} hearing aid group. Nine children who used hearing aids demonstrated hearing levels of 101-110 dB HL at two of the three frequencies; mean PTA was 104 dB HL. These children were designated as the PTA_{104} group. Summary information for all five groups (cochlear implant users at the early interval, the same children at the late interval, tactile aid users, PTA_{93} hearing aid users, and PTA_{104} hearing aid users) is shown in Table 1.

Table 1: Subject Characteristics

	Cochlear Implant Users	Tactile Aid Users	Hearing Aid Users	
			PTA_{93}	PTA_{104}
Subjects	*n* = 9	n = 9	*n* = 9	*n* = 9
Mean unaided pure tone average, dB HL	111	>110	93	104
Mean age at onset of deafness, years	0.41	0.53	0.13	0.14
Mean age fit with sensory aid, years	5.38	5.78	1.81	1.04
Mean years deaf	4.97	5.24	1.68	.90
Mean age at time of testing, years				
Early interval	5.67	n/a	n/a	n/a
Late interval	7.51	7.61	7.78	7.47
Mean duration of device use, years	2.13	1.83	5.97	6.42

Speech Materials and Analysis

Consonant production was elicited by administration of the Sounds-in-Words Subtest of the *Goldman-Fristoe Test of Articulation* (Goldman & Fristoe, 1972; hereinafter GFTA), a standard instrument used for assessing articulation disorders and articulation development in children. In addition to its widespread use with various pediatric clinical populations (including children with hearing impairments; see Osberger, Robbins, Lybolt, Kent, & Peters, 1986), the GFTA also has the advantage of using actual words and of probing sound segments in different word-positions (word-initial, -medial, and -final). The test consists of 44 words containing all English consonants (except [ʒ]) that can occur in word-initial and most in word-medial and word-final position. Stimulus materials consist of colored drawings on card stock.

502

The GFTA is part of a battery of speech perception, speech production, and language tests routinely administered to all participants (including control subjects) in sensory aid studies in the Department of Otolaryngology–Head and Neck Surgery at the Indiana University School of Medicine.

All productions elicited in the GFTA were recorded on audiotape and later transcribed phonetically by a speech-language pathologist with experience hearing and transcribing the speech of deaf children. Transcriptions were entered into the Logical International Phonetics Programs (LIPP 1.40; Intelligent Hearing Systems, Miami, Florida), a computerized transcription and phonetic/phonological analysis program. With this program, a user enters a phonetically transcribed target form and a segment-aligned phonetic transcription of the corresponding production. LIPP subsequently performs a number of segment-by-segment analyses, including analyses of feature and segment correctness. For the present study, the program generated percent correct production scores for the segments, features, and feature classes listed in Table 2.[4]

Table 2: LIPP Output

Feature Class	Feature	Segments
Voicing	Voiced	b d g v ð z dʒ m n w l
	Voiceless	p t k f θ s ʃ tʃ h
Place of Articulation	Bilabial	p b m
	Dental	θ ð
	Alveolar	t d n l s z
	Velar	k g
Manner of Articulation	Stop	p b t d k g
	Fricative	f v θ ð s z ʃ h
	Affricate	tʃ dʒ
	Nasal	m n
	Glide	w l

---

[4] To facilitate analysis, LIPP was programmed with the following settings: (1) only singletons (not clusters) were included in the analysis; (2) four places of articulation were included in the analysis: bilabial, interdental, alveolar, and velar; (3) glides were limited to [w] and [l] and were not included in the Correct Manner score; (3) the velar nasal [ŋ] was not included in the

Results

Development of Consonant Feature Production in Cochlear Implant Users

Figure 1 shows mean percent correct scores for voicing, place of articulation, and manner of articulation for cochlear implant users at both the early interval and the late interval. As this figure shows, mean scores (collapsed across word positions) at the late interval were higher for all three feature classes than at the early interval.

Insert Figure 1 about here

To determine if the improvements demonstrated by the cochlear implant users were significant, a separate paired *t* test was computed for each feature class (voicing, place or articulation, manner of articulation) with interval as the independent variable and percent correct score as the dependent variable. For these cochlear implant users, percent correct scores for voicing ranged from 0 to 85% ($M = 30.6$, $SD = 22.9$) at the early interval and from 6 to 94% ($M = 52.5$, $SD = 20.6$) at the late interval. A paired t test showed the difference in mean correct percent between the two intervals to represent a significant improvement in production of correct voicing ($t = -6.39$, $p < .001$). Similarly, percent correct scores for place of articulation ranged from 0 to 55% ($M = 21.1$, $SD = 19.2$) at the early interval and from 6 to 94% ($M = 47.8$, $SD = 23.1$) at the late interval, a significant increase ($t = -7.62$, $p < .0001$). Finally, percent correct scores for manner of articulation ranged from 0 to 55% ($M = 19.1$, $SD = 17.6$) at the early interval and from 0 to 94% ($M = 40.3$, $SD = 24.0$) at the late interval; this increase was also significant ($t = -5.78$, $p < .0001$). Table 3 summarizes the results of the paired *t* tests.

Table 3: Paired *t* test results: Early vs. Late Interval Percent Correct Production of Voicing, Place, and Manner Features by Cochlear Implant Users

Feature Class	Percent correct Early Interval, *M*	 Late Interval, *M*	Difference (Late - Early)	*SEM* of paired differences	*t*
Voicing	30.6	52.5	21.9	3.4	6.39*
Place	21.1	47.8	26.7	3.5	7.62**
Manner	19.1	40.3	21.2	3.7	5.78*

$^*p < .001$; $^{**}p < .0001$

For the nine cochlear implant users, then, production improved significantly from the early interval to the late interval for all three feature classes examined in this study: voicing, place of articulation, and manner of articulation.

---

analysis. Additionally, production of target initial [tʃ] was assessed from the word *chicken*, rather than from *church*, the former having been found to be more familiar to children.

504

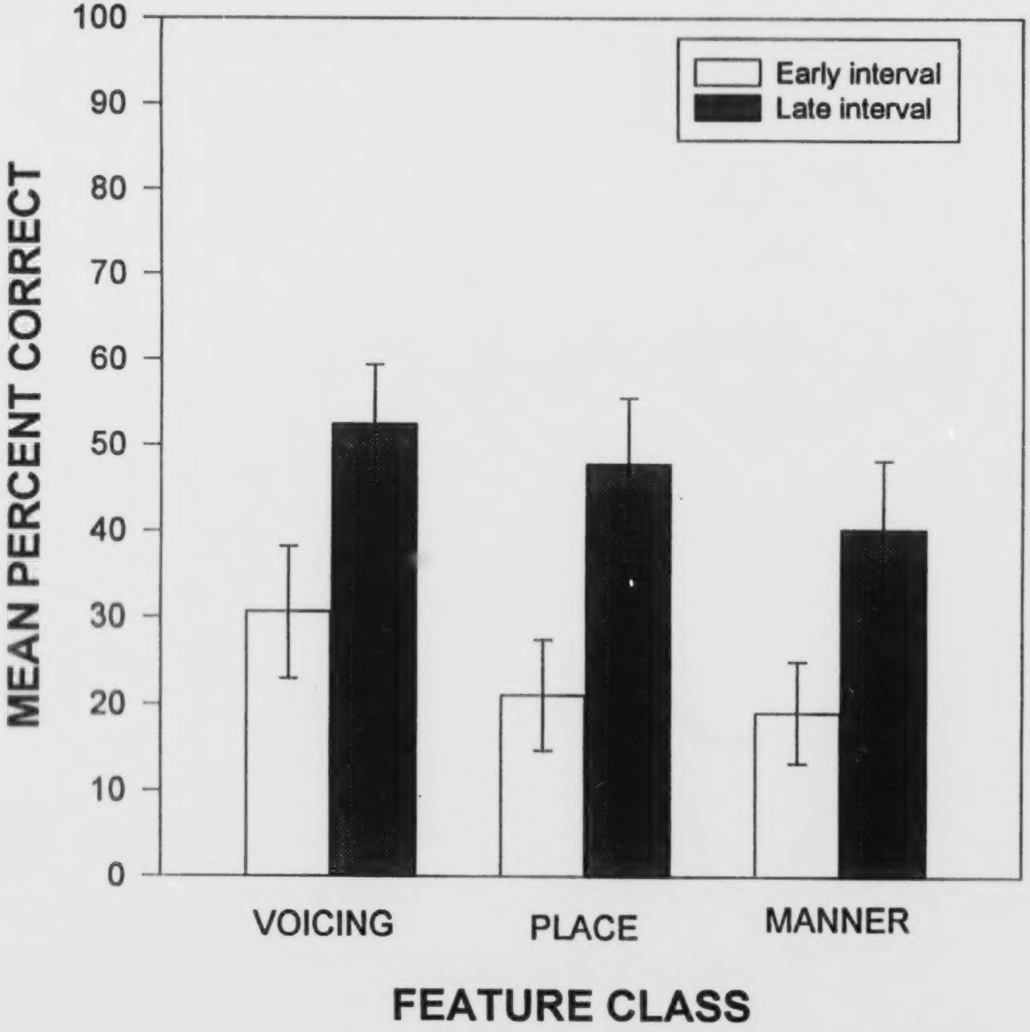


Figure 1: Cochlear implant users' mean percent correct production by feature class and interval. Error bars indicate standard errors of means.

Sensory Aid Effects on Consonant Feature Production

Figure 2 shows mean percent correct scores achieved by PTA_{93} hearing aid users, cochlear implant users at the late interval, PTA_{104} hearing aid users, and tactile aid users for the feature classes voicing, place of articulation, and manner of articulation. As this figure shows, scores for PTA_{93} hearing aid users were consistently higher than scores from the other device groups, but cochlear implant users also consistently outperformed both PTA_{104} hearing aid users and tactile aid users on correct production of all three types of features.

Insert Figure 2 about here

Across word positions, percent correct scores for PTA_{93} hearing aid users ranged from 18 to 100% ($M = 77.9$, $SD = 13.6$) for voicing, from 22 to 100% ($M = 71.8$, $SD = 15.9$) for place of articulation, and from 17 to 95% ($M = 64.1$, $SD = 17.9$) for manner of articulation. For cochlear implant users at the Late interval, scores ranged from 6 to 94% ($M = 52.5$, $SD = 20.6$) for voicing, from 6 to 94% ($M = 47.8$, $SD = 23.1$) for place of articulation, and from 0 to 94% ($M = 40.3$, $SD = 24.0$) for manner of articulation. For PTA_{104} hearing aid users, percent correct scores ranged from 0 to 80% ($M = 40.9$, $SD = 20.0$) for voicing, from 0 to 78% ($M = 35.1$, $SD = 22.1$) for place of articulation, and from 0 to 65% ($M = 32.2$, $SD = 20.3$) for manner of articulation. Finally, for tactile aid users, percent correct scores ranged from 0 to 75% ($M = 34.2$, $SD = 17.4$) for voicing, from 0 to 95% ($M = 33.6$, $SD = 25.3$) for place of articulation, and from 0 to 90% ($M = 31.2$, $SD = 25.2$) for manner of articulation.

To determine the statistical significance of differences among the device groups, a separate one-way ANOVA was computed for each feature class, using device group as the independent variable and feature class score as the dependent variable. These revealed a significant main effect of device on percent correct production of voicing ($p < .001$), place ($p < .001$), and manner ($p = .001$) features. Post-hoc analyses (Student-Newman-Keuls method) revealed PTA_{93} hearing aid users' performance to be significantly higher than that of the three other groups on all three feature classes. There were no significant differences between the performance of the cochlear implant users and the PTA_{104} hearing aid and tactile aid users, except that performance by the cochlear implant users on voicing feature production was significantly higher ($p < .05$) than that of the tactile aid users.

Word-Position Effects on Consonant Feature Production

A two-way RMANOVA was computed to determine the effects of device group and word position (word-initial, -medial, and -final) on feature class scores. This indicated a significant main effect of word position on feature production scores [$F_{2,64} = 62.86$, $p < .0001$]. Figure 3 displays percent feature production scores as a function of device group and word position.

Insert Figure 3 about here

As Figure 3 shows, there was a consistent pattern across devices, such that correct feature production was highest in initial position, followed by medial position, with features in final position produced least correctly. All device groups showed a significant ($p < .05$) difference between initial and final position.

507

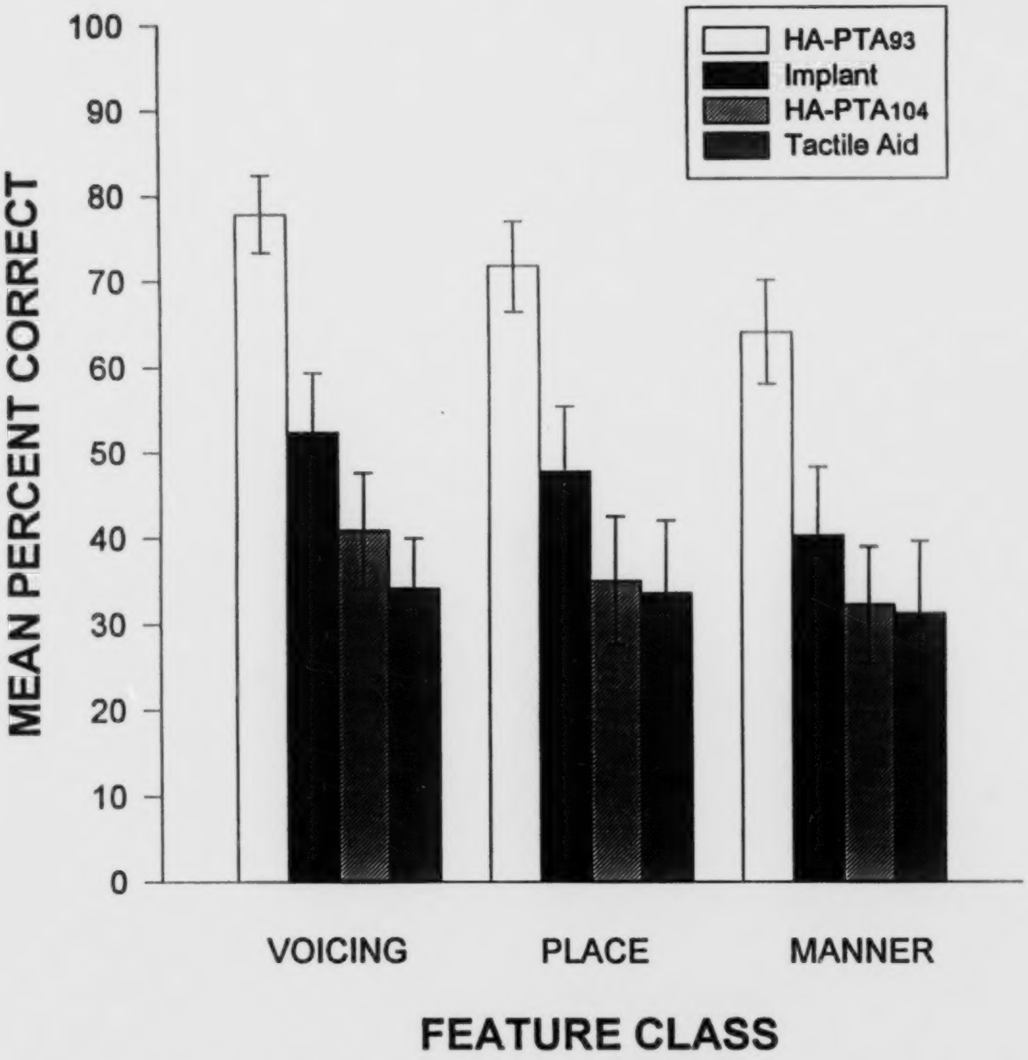


Figure 2: Mean percent correct production by feature class and device. Error bars indicate standard errors of means.

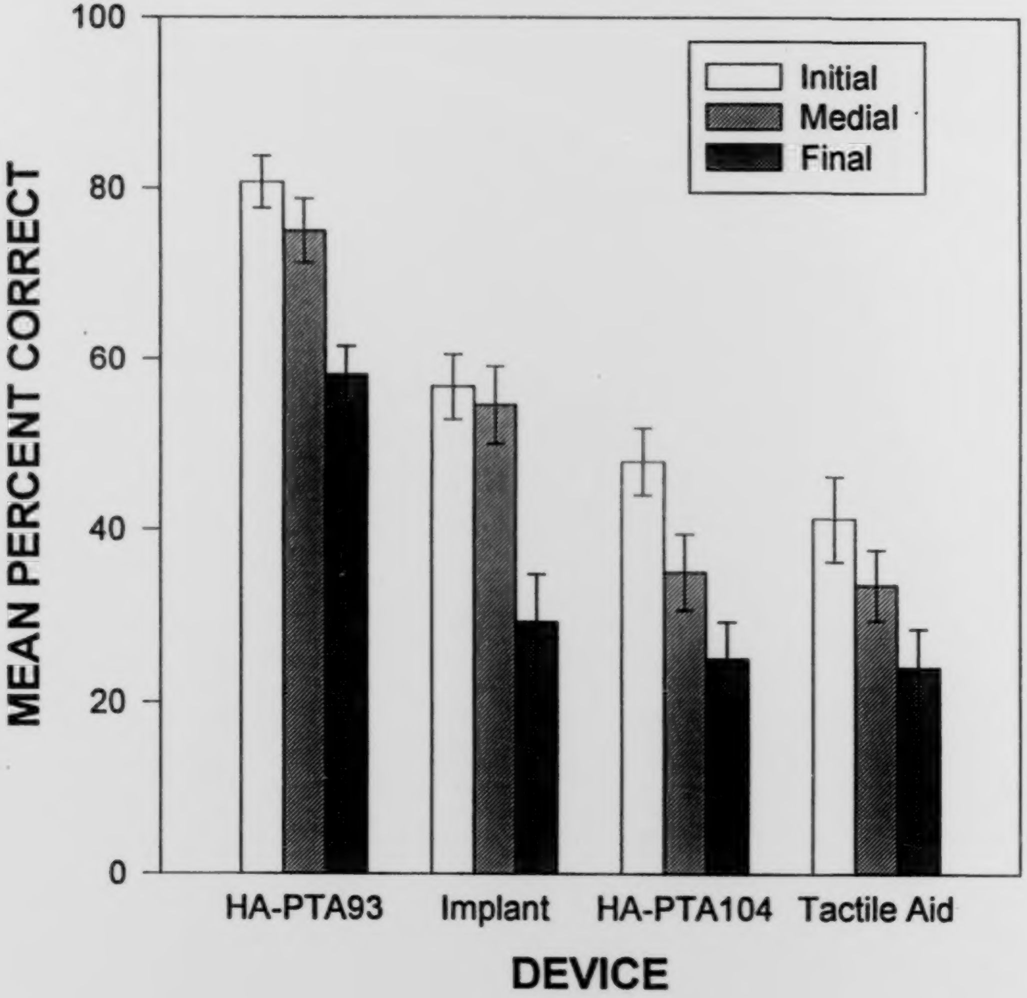


Figure 3: Mean percent correct production by word position and device. Error bars indicate standard errors of means.

Additionally, both PTA_{93} hearing aid users and cochlear implant users displayed a significant difference ($p < .05$) between medial and final position but not between initial and medial position. Neither PTA_{104} hearing aid nor tactile aid users showed a significant difference between medial and final position. Finally, PTA_{104} hearing aid users showed a significant difference between initial and medial position, but tactile aid users did not. These results are summarized in Table 4.

Table 4: Significant Differences ($p < .05$) in Scores for Device Groups as a Function of Word Position Opposition

Word Position Opposition	PTA_{93} Hearing Aid	Cochlear Implant	PTA_{104} Hearing Aid	Tactile Aid
Initial vs. Medial	No	No	Yes	No
Initial vs. Final	Yes	Yes	Yes	Yes
Medial vs. Final	Yes	Yes	No	No

As Table 4 shows, PTA_{93} hearing aid users and cochlear implant users displayed a common pattern in their differentiation of word positions. In addition, both displayed common differences from the patterns shown by PTA_{104} hearing aid and tactile aid users.

The two-way RMANOVA also indicated a significant main effect of device type on percent correct feature production [$F_{3,32} = 6.55$, $p < .01$]. Figure 4 shows percent correct feature class production as a function of word position and device group.

Insert Figure 4 about here

As Figure 4 shows, there was a consistent pattern of scores across the three word positions, such that PTA_{93} hearing aid users scored highest, followed by cochlear implant users, and then PTA_{104} hearing aid and tactile aid users. In initial and medial positions, differences in scores between the PTA_{93} hearing aid users and the PTA_{104} hearing aid and tactile aid users were significant ($p < .05$); importantly, scores for cochlear implants users and PTA_{93} hearing aid users did not differ significantly. In final position, where scores were lowest regardless of device, there were no significant differences due to type of device. Finally, there were no significant interaction effects of device type and word position [$F_{6,64} = 1.79$, $p = .1149$ (n.s.)].

Discussion

The results of this study show that users of cochlear implants improved significantly on their correct production of voicing, place of articulation, and manner of articulation features from the early interval to the late interval. This finding indicates that cochlear implants promote the development of appropriate nonimitative speech production in children with profound hearing impairments.

When the production of cochlear implant users after approximately two years was compared with that of PTA_{93} hearing aid users, PTA_{104} hearing aid users, and tactile aid users of the same chronological age, results were consistent with previous studies (e.g., Sehgal et al., 1998). That is, PTA_{93} hearing aid users demonstrated significantly better production of consonant features than did other device groups.

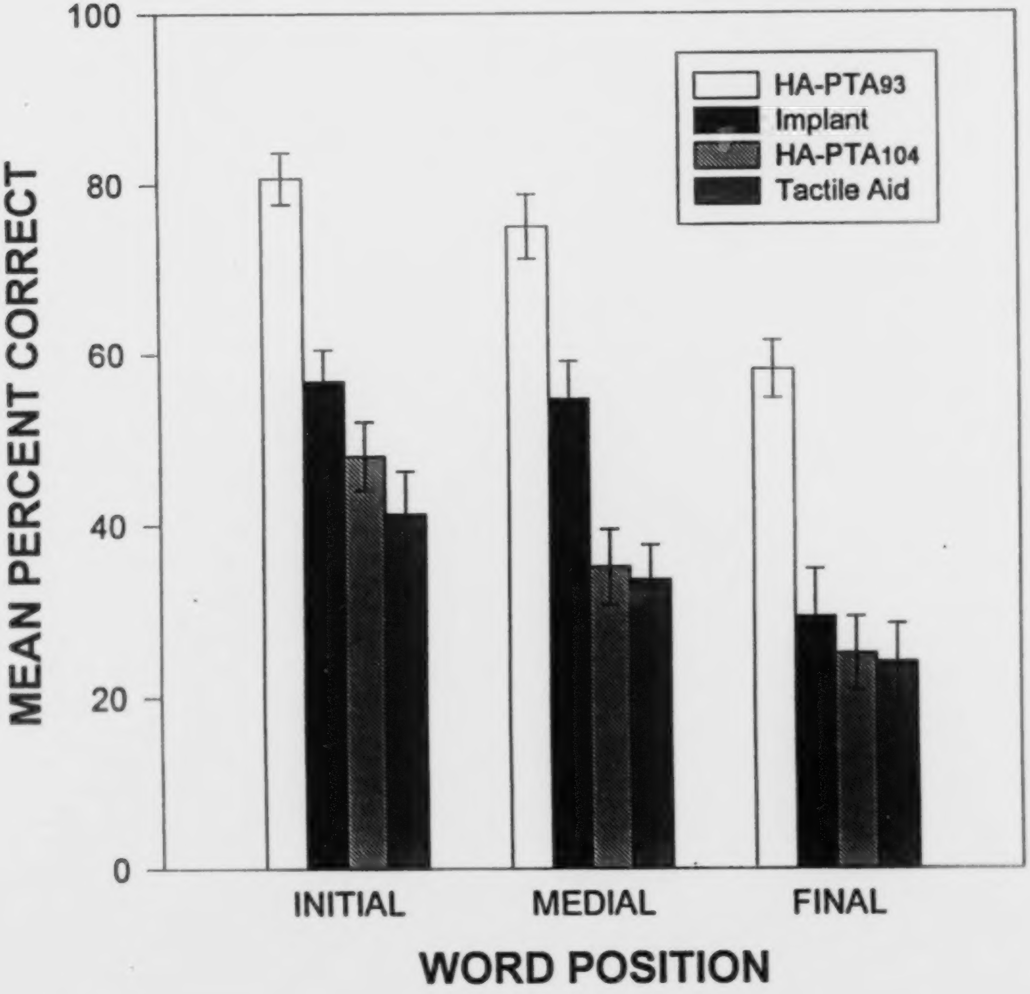


Figure 4: Mean percent correct production by device and word position. Error bars indicate standard errors of means.

However, there was a trend for the cochlear implant users to produce consonant features more correctly than either PTA$_{104}$ hearing aid users or tactile aid users. Previous research has shown that consistent speech perception benefits of cochlear implants frequently are apparent only after approximately two years of device use. If production skills lag behind perception skills, this may account for the general lack of significant differences between cochlear implant users and PTA$_{104}$ hearing aid and tactile aid users, as well as for the fact that PTA$_{93}$ hearing aid users scored significantly higher than cochlear implant users at the late interval. Moreover, because of the additional task demands, it is probable that the type of nonimitative speech production elicited in this study is considerably more difficult than imitative speech production and may develop at a later age (or with more device experience, especially if children are implanted earlier). Thus, it is possible that only after continued experience with a cochlear implant beyond the stage reported here will differences among cochlear implant, hearing aid, and tactile aid users be significant.

This same observation is relevant for interpretation of the results stemming from comparison of production in different word positions. There were consistent and generally significant differences in scores for different word positions across device types, and the hierarchy of production ability (initial > medial > final) is consistent with findings for children with normal hearing (e.g., Ingram, 1989; Templin, 1957). Word position effects on consonant production thus apply equally to both hearing-impaired and normal-hearing children. In comparing differences between word positions, it was also found that cochlear implant users and PTA$_{93}$ HA users demonstrated a pattern not shown by the tactile aid users and PTA$_{104}$ HA users. Specifically, cochlear implant users and PTA$_{93}$ HA users were alike in showing significant differences in performance between initial and final positions and between medial and final positions, but not between initial and final positions. The pattern of differentiating word position exhibited by PTA$_{93}$ hearing aid users and cochlear implant users is furthermore consistent with that demonstrated by children with normal hearing (e.g., Templin, 1957; Wellman et al., 1936). The difference in performance may be due to differences in the acquisition of English syllabification. PTA$_{93}$ hearing aid users and users of cochlear implant users appear to be aware that in English, word-medial singleton consonants are generally syllabified with the following syllable (as onset). This might account for the lack of significant differences in performance between initial and medial positions, as well as for the significant differences between initial and final position and between medial and final position. For PTA104 hearing aid users and users of tactile aids, however, medial consonants may be either ambisyllabic or syllabified with the preceding syllable. This might account for the lack of significant difference in performance between medial and final position. Continued experience with cochlear implants may illuminate further differences in performance among users of various sensory aids, in particular differences between users of cochlear implants and users of hearing aids with considerable residual hearing (e.g., the present PTA$_{93}$ hearing aid users).

There is thus an increasing need to examine linguistically relevant speech production among users of cochlear implants and other sensory aids. It is to be expected that as these children mature, their awareness of sound-to-meaning correspondences will become more acute. Furthermore, spoken language acquisition is not a matter of pure imitation or of producing consonants only at the beginnings of words. As children with profound hearing impairment and various sensory aids continue to mature, develop, and expand their linguistic abilities, and as new processing strategies (e.g., CIS, SPEAK) become available, it is incumbent on researchers to expand their own assessment repertoires as well.

References

Berliner KI, Eisenberg LS, House WF. (1985). The cochlear implant: An auditory prosthesis for the profoundly deaf child [Preface]. *Ear and Hearing, 6* (Suppl.), 4S-5S.

Ertmer DJ, Kirk KI, Sehgal ST, Riley AI, MJ Osberger. (1997). A comparison of vowel production by children with multichannel cochlear implants or tactile aids: Perceptual evidence. *Ear and Hearing*, **18**, 307-315.

Franklin D. (1991). *Temporary users manual for the Tactaid 7*. Somerville, MA; Audiological Engineering Corporation.

Geers A, Moog J. (1994). Description of the CID sensory aids study. *Volta Review*, **96(5)**, 1-11.

Goldman R, Fristoe M. (1972). *The Goldman-Fristoe test of articulation*. Circle Pines, MN: American Guidance Service.

Ingram D. (1989). *Phonological disability in children*. (Second edition). London: Cole & Whurr.

Kirk KI, Diefendorf E, Riley A, Osberger MJ. (1995). Consonant production by children with multichannel cochlear implants or hearing aids. In AS Uziel & M Mondain (Eds.), *Cochlear implants in children* (*Advances in Oto-Rhino-Laryngology*, **50**) (pp. 154-159). Basel, Switzerland: S. Karger.

Leder S, Spitzer J. (1990). A perceptual evaluation of the speech of adventitiously deaf adults males. *Ear and Hearing*, **11**, 169-175.

Ling D. (1976). *Speech and the hearing-impaired child: Theory and practice*. Washington, DC: Alexander Graham Bell Association.

McGarr N. (1983). The intelligibility of deaf speech to experienced and inexperienced listeners. *Journal of Speech and Hearing Research*, **26**, 451-459.

Miyamoto RT, Osberger MJ, Todd SL, Robbins AM. (1994). Speech production of children with multichannel cochlear implants. In IJ Hochmair-Desoyer & ES Hochmair (Eds.), *Advances in cochlear implants* (pp. 408-502). Vienna, Austria: Manz.

Osberger MJ, Maso M, Sam L. (1993). Speech intelligibility of children with cochlear implants, tactile aids, or hearing aids. *Journal of Speech and Hearing Research*, **36**, 186-203.

Osberger MJ, Robbins AM, Lybolt J, Kent RD, Peters J. (1986). Speech evaluation. In MJ Osberger (Ed.), *Language and learning skills of hearing-impaired students* (ASHA Monographs Number 23) (pp. 24-31). Rockville, MD: American Speech-Language-Hearing Association.

Patrick JF, Clark G. (1991). The Nucleus 22-channel cochlear implant system. *Ear and Hearing*, **12** (Suppl.), 10S-14S.

Sehgal ST, Kirk KI, Svirsky M, Ertmer DJ, Osberger MJ. (1998). Imitative consonant feature production by children with multichannel sensory aids. *Ear and Hearing*, **19**, 72-84.

Templin MC. (1957). *Certain language skills in children: Their development and interrelationships*. Minneapolis: The University of Minnesota Press.

Tobey E, Angelette S, Murchinson C, Nicosia J, Sprague S, Staller SJ, Brimacombe J, Beiter AL. (1991). Speech production performance in children with multichannel cochlear implants. *American Journal of Otology*, **12** (Suppl.), 165-173.

Tobey E, Geers A, Brenner C. (1994). Speech production results: Speech feature acquisition. *Volta Review*, **96**(5), 109-129.

Tobey EA, Hasenstab MS. (1991). Effects of a Nucleus multichannel cochlear implant upon speech production in children. *Ear and Hearing*, **4** (Suppl.), 48S-54S.

Tobey EA, Pancamo S, Staller SJ, Brimacombe, Beiter AL. (1991). Consonant production in children receiving a multichannel cochlear implant. *Ear and Hearing*, **12**, 23-31.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 21 (1996-1997)
Indiana University

Lexical Competition and Reduction in Speech: A Preliminary Report[1]

Richard Wright

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405

---

[1] This work supported by NIH-NIDCD Training Grant DC0012 to Indiana University Bloomington. The author is grateful to Stefan Frisch, Ann Bradlow, and Gina Torretta for helpful comments during preparation of this manuscript.

Lexical Competition and Reduction in Speech: A Preliminary Report

Abstract. Lindblom (1990) among others has proposed that talkers accommodate listeners' communicative needs by controlling the degree of reduction (hyper- and hypo-articulation) in different contextual conditions, thereby maintaining sufficient intelligibility of words across a variety of contexts. Lindblom's proposal predicts that lexical factors that affect intelligibility of a word will affect the hypo- and hyper-articulation of words. Based on factors in lexical competition such as usage frequency and similarity-neighborhood density, previous research has characterized words as "easy" or "hard" to identify. This study examines the degree of centralization of vowels (a well known feature of reduction or hypo-articulation) in 34 "easy" and 34 "hard" monosyllabic (CVC) words of equal familiarity spoken in isolation by 10 talkers. Measurements of the first two vowel formants (F1, F2) were made at the point of maximal displacement in the vowel (excluding the initial and final 50 ms of the vowel). Centralization is measured by calculating the Euclidean distance from the center of a talker's F1-F2 vowel space. Three results emerge: 1) overall "easy" words were significantly more centralized than "hard" words, 2) peripheral vowels, such as /i/, /a/, /u/, showed the greatest effect, and 3) there was considerable between talker variability in the magnitude of the difference between vowels in "easy" and "hard" words. The results are interesting because they demonstrate that the talker takes into account a wider variety of sources of possible noise and information than previously thought. These results have implications for both diachronic and synchronic processes that involve reduction in linguistics. They have further implications for applications in speech recognition that model variation in spoken language.

Introduction

Understanding sources of variability in spoken language is one of the most important challenges that face speech researchers today. It is an issue that spans several fields such as linguistic phonetics with its interest in language specific vs. universal traits of spoken language; historical linguistics and sociolinguistics, which study patterns of sound change; and engineering with its interest in improving automatic speech recognition systems. Traditionally, variability had been treated as noise to be controlled or ignored in studying spoken language; however, there has recently been an increasing interest in exploring lawful variability in spoken language. Variability is a pervasive characteristic of spoken language that is introduced at nearly every level of an utterance. It goes well beyond the frequently noted physiologically based factors such as differences in larynx structure and vocal tract length. For example, in a study that examined a large number of talkers and utterances using the TIMIT speech database, Byrd (1994) found that sex and to a lesser degree dialect differences resulted in between-talker differences in the degree of reduction along a number of dimensions including speech rate, stop release, flapping, and quantity of central vowels. In addition to inter-talker differences which might be viewed as talker or group specific constants, there are many forces that act on spoken language that may change the way a word is pronounced from one utterance to the next by the same talker. Reduced and clear speech processes represent a significant source of within talker variability, much of which has been attributed to talkers varying their pronunciation to accommodate the communicative needs of the listener (e.g., Anderson, Bard, Sotillo, Newlands, & Doherty-Sneddon, 1997; Bolinger, 1963; Lieberman, 1963; Lindblom, 1990). In these studies talkers have been shown to produce more reduced speech when contextual information within the utterance or in the environment can aid the listener in recognizing what is said, and to produce more careful speech in when the talker is aware of conditions that may impede the listener's ability to understand what is said. When words are isolated from their spoken contexts, the speech produced under reducing conditions has a low intelligibility and the more careful speech has a high intelligibility. Although most studies have concentrated on contextual and environmental factors, relatively few have examined how word specific characteristics might affect production strategies. The purpose of this study is to examine the degree to which factors in lexical competition that are known to affect intelligibility of individual words influence the carefulness which talkers produces words.

Reduction and Sources of Information

Although many studies have assumed reduction as a constant, Lindblom (1990) has proposed a more explicit model of the interaction between the forces that shape both reduced and clear speech. In his model, speech motor control is output-oriented and plastic. In this view, reduced speech and clear speech lie along a continuum of contextually determined variability. In a communicative context there is pressure on the system to maintain sufficient information in the signal for the listener to recover the intended message. As factors decrease the probability that a listener will be able to recover the message, in the talker's estimation, output constraints become more severe and the talker is forced to produce clearer speech (i.e., exaggerate contrast among lexical items), which Lindblom calls "hyper-speech." As output constraints become less severe, the system tends towards economy of effort becoming more system-oriented, resulting reduced speech ("hypo-speech"). In this view, speech perception involves discrimination among stored items—lexical access is a function of distinctiveness not invariance. Thus reduction and hyper-articulation will occur along dimensions that will increase or decrease the perceptual distance among lexical items. Reduction in speech can be measured in a number of ways; some of the better known characteristics are shortening of vowels, increased flapping, increased consonant and vowel deletion, and vowel centralization.

A commonly cited example of the talker modifying pronunciation to accommodate the hearer's needs is the Lombard effect (Lane & Tranel, 1971; Lane, Tranel, & Sisson, 1970; Lombard, 1911), in which the talker produces more careful and higher intensity speech in response to increased environmental noise. Similarly, in speech directed towards the hard of hearing there is less phonological reduction and an overall slower rate than is observed in conversational speech (Picheny, Durlach, & Braida, 1986; Uchanski, Choi, Braida, Reed, & Durlach, 1996). More careful productions have also been observed when talkers introduce unexpected or novel information into the discourse (Bolinger, 1963; Chafe, 1974; Hawkins & Warren, 1994). On the other hand, reduction occurs when the talker estimates that a listener will have little difficulty in identifying a word. For example, Lieberman (1963) found that words that are highly predictable from sentential context are more reduced than identical words produced in an equivalent sentence that does little to narrow the field of lexical candidates. When presented (without the rest of the sentential context) as stimuli to listeners, the words from the highly predictive contexts were less intelligible than the words from the less redundant contexts. Similar results have been found for a variety of predictive contexts (Bard & Anderson, 1983; Fowler & Housum, 1987; Hunnicutt, 1985).

Talkers are sensitive to a wide variety of sources of information not only from the auditory mode but also from the visual mode (and one would predict the haptic mode as well). For example, Anderson, Bard, Sotillo, Newlands, & Doherty-Sneddon (1997) found that a talker's performance reflects the listener's access to visual information from the talker's face. Visual information in the talker's face can aid the hearer enormously in recovering the spoken message: the equivalent of as much as an 18 dB gain by some estimations (Sumby & Pollack, 1954). Anderson et al. (1997) found that tokens in spoken language when the talker estimated that listener had access to visual information were significantly degraded. In later intelligibility tests, the reduction of the auditory signal was offset only if subjects were given access to the accompanying video stimuli in addition to the audio stimuli. An interesting point that emerged from the study was that talkers maintained only a loose model of the information available to the listener; rather than tracking the listener's use of visual information from moment to moment, the talker adjusts the carefulness of speech in a more global fashion basing the model on talker internal conditions. That is, talkers did not pay attention to whether or not the listener was actually looking at his/her face, but rather to whether or not the listener had the ability to look.

Lexical Competition and Intelligibility

In addition to the message internal and environmental contextual factors that affect a word's intelligibility, there are lexical factors that may increase or decrease the probability of a talker identifying a word correctly. Word frequency is perhaps the best known lexical trait that may affect a word's intelligibility. It is also a proposed factor in overall word shortening and acceleration of reduction processes (Balota, Boland, & Shields, 1989, Bybee, 1994; Zipf, 1935). However, overall word frequency alone has proven to be a rather poor predictor of intelligibility (Luce, 1986; Pisoni, Nusbaum, Luce, & Slowiaczek, 1985) and has proven an unreliable predictor of reduction. Rather, word identification must be viewed in the context of lexical competition for the role of frequency in intelligibility to be clearly seen. In his dissertation, Luce (1986) studied patterns of auditory word confusion and found that a word's intelligibility is affected by two lexical factors: 1) *neighborhood density*: the number of phonologically similar words in the language, and 2) *relative frequency*: the frequency of the target word relative to its nearest phonological neighbors. In calculating nearest neighbors, Luce found a reasonably close match between his confusion matrices and neighborhood density determined using the *single phoneme substitution* method (Greenberg and Jenkins, 1964) in which all words that differ from the target word by a single phoneme are considered nearest neighbors. Luce proposed the Neighborhood Activation Model (NAM) in which the number of similar competitors that a word has and usage frequency have inhibitory and excitatory effects in lexical access and competition. In this model, while frequency will determine the probability of a word beating out its neighbors, a word with few neighbors is likely to be identified even if its usage frequency is low. Based on intelligibility properties, words from high density neighborhoods and with low relative frequency have been termed "hard" and those from low density similarity neighborhoods and with high relative frequency have been termed "easy" (Luce, 1986; Pisoni, Nusbaum, Luce, & Slowiaczek, 1985).

A preliminary study of the effect of neighborhood density on voice onset time (VOT), one of the main indicators of stop consonant voicing, was conducted by Goldinger and Summers (1989). In their study, they had talkers read minimal pairs of CVC words in which the word-initial stop consonant was either a voiced or a voiceless stop. The pairs were chosen so that both were from sparse neighborhoods or both were from dense neighborhoods. Each talker read each pair of words four times. Overall there was a greater difference in VOT between voiced-voiceless pairs from dense neighborhoods than in voiced-voiceless pairs from sparse neighborhoods. That is, the voicing contrast that is cued in part by VOT was more exaggerated in minimal pairs from dense neighborhoods than in minimal pairs from sparse neighborhoods. Moreover, across repetitions the difference in VOT between the dense pairs increased dramatically while the VOT difference between the sparse pairs increased slightly. The study was flawed because the method of presenting the words in minimal pairs attracts the talkers attention to the contrast being studied and generally results in an exaggeration of the contrast. Nevertheless, the fact that the sparse and dense neighborhoods predicted differences in talkers' behavior over repetitions is a preliminary indication that neighborhood density is a factor in variability of spoken language. Interestingly, the fact that all the words in the study showed the effect indicates that neighborhood density appears to have a global word level effect, at least on monosyllabic words, because the majority of lexical competitors would not have been confusable on the first phoneme. That is, although some of the words may have had neighbors that were close because they differed only in the initial phoneme, the single phoneme substitution method that the authors employed to calculate neighborhood density implies that roughly two thirds of the neighbors are based on the second or third phonemes in the words. The apparent whole word effect implies that talkers have only a loose sense of sources of neighborhood density rather than making fine adjustments. This last point is similar to the observation in Anderson et al. that talkers maintain only a crude estimation of what visual information listeners may be using.

The current study was designed with these preliminary VOT findings in mind. Because of the interaction between neighborhood density and relative frequency in the intelligibility studies mentioned above, it was decided that the first place to look for an effect was in "easy" vs. "hard" words. This choice was further driven by the existence of a publicly available prerecorded database of "easy" and "hard" tokens. One advantage of using a database is that a relatively large number of talkers can be studied, ten in this case. A second advantage is that the words of interest are well studied in independent research; tokens in the database have been measured for intelligibility and have been pre-coded for lexical characteristics. There are disadvantages to working with databases. The biggest one is lack of control the experimenter has over the recording conditions and over the choice of recorded material. In this case it would have been ideal to be able to have access to more detailed demographic information about the talkers and to be able to make follow-up recordings. Despite these shortcomings, this database and others like it provide researchers with rich tools for testing hypotheses about spoken language. It is predicted in this study that factors in lexical competition should affect the degree of reduction and in production just as contextual factors do. That is, "easy" words should show a greater degree of reduction than "hard" words. Because differences in vowel centralization have been shown to differentiate clear from casual speech (e.g., Byrd, 1994; Lindblom, 1990; Lindblom & Moon, 1994) and because increased vowel dispersion is an established correlate of intelligibility (Bond & Moore, 1994; Bradlow, Torretta, & Pisoni, 1996; Picheny, Durlach, & Braida, 1985), "hard" words are predicted to have vowels that are more dispersed (i.e., less centralized) than "easy" words. Increased dispersion is characterized by an expanded vowel space or by overall increased acoustic distances between vowels from different categories.

Method

Recording Materials

Tokens in the study were all monosyllabic CVC words drawn from a prerecorded database (see Torretta, 1995, for a detailed description). All words were of equally high familiarity, being between 6.8 and 7.0 on the 1-7 point Hoosier Mental Lexicon scale (Nusbaum, Pisoni, & Davis, 1984), but varied crucially in their similarity neighborhood density and in relative usage frequency. One set of words, termed "easy," came from sparse similarity neighborhoods and had usage frequencies that were high relative to their neighbors. A second set of words, termed "hard," came from dense similarity neighborhoods and had usage frequencies that were low relative to their neighbors. The words were chosen to provide a balanced segmental context for the vowel; consonantal contexts that could result in vowel coloring were avoided while maintaining similar contexts in easy and hard words. For example, postvocalic /r/ and /l/ were avoided altogether for both types of words, and nasal codas, when unavoidable, were balanced in both sets. The full set of tokens is listed in Appendix 1. Overall, there are 34 "easy" and 34 "hard" words spoken by 10 talkers (5 male and 5 female) of American English resulting in 680 tokens total. The sound files in the database are digital recordings of monosyllabic words presented singly in pseudo-random order on a CRT monitor and read once in isolation. The talkers were instructed to say each word at a "medium" rate. The utterances were antialias filtered and digitized directly to disk at 22050 Hz (see Torretta, 1995, for a detailed description).

Measurement

As the file names in the database included lexical neighborhood information, the files were renamed and randomized prior to measurement. The first and second formants (F1 and F2) of each vowel were measured at the point of maximal displacement on all 680 tokens. The initial and final 50 ms of the vowel were excluded to minimize the effect of flanking consonants on the measurement. The point of maximal displacement occurs when F1 and F2 are the most characteristic for that particular vowel. For example, for the vowel /i/ it is the point where F1 is lowest and F2 is highest and for /a/ it is the point where F1 is highest and F2 is lowest. Where F1 and F2 were not in agreement, F1 was taken as the point of reference and F2 was measured at that point. This measure is equivalent to what has typically been described in the literature as the "steady state" but takes into account the dynamic character of vowels which often results in the absence of a clear steady state. For diphthongs the measure was taken in the primary portion of the vowel (e.g., for /aj/ the measure was for the /a/ portion). Formant values were measured from a twelfth order LPC with a 25 ms window overlaid on a simultaneous 512 point FFT. A wideband spectrogram was used to locate the measurement point and for reference during the formant measures. The formant values were converted into the Bark scale (an auditory transform) using the formula given in (1) (Zwicker & Terhardt, 1980) where Z is bark and *f* is frequency in Hertz.

$$Z=\left[\frac{26.81f}{(1960+f)}\right]-0.53 \qquad (1)$$

The degree of dispersion was measured using a technique applied in Bradlow, Torretta, & Pisoni's (1996) study of talker intelligibility: the Euclidean distance from the center of a talker's F1 by F2 vowel space. Using this measure, Bradlow et al. found that one of the best correlates of talker intelligibility was the degree of vowel dispersion. Differences in dispersion were submitted to an analysis of variance with dispersion as the dependent variable and lexical category (easy/hard), vowel category, and talker as the independent variables.

Results and Discussion

Overall, The hypothesis was borne out; there was a reliable effect of lexical category ("easy" vs. "hard") on dispersion (the Euclidean distance from the center of the vowel space). With an alpha level of .01 the analysis revealed a significant main effect of lexical category, $F(1,480) = 130.92$, $p < .0001$. There was also a significant interaction between lexical category and vowel type, $F(9,480) = 15.22$, $p < .0001$. Figure 1 shows the overall dispersion in Bark (vertical axis) in the height of the bars for "easy" vs. "hard" words collapsing across talker and vowel type. There is a clear difference in the degree of dispersion with the vowels from "hard" words being more dispersed on average than the vowels from "easy" words.

Insert Figure 1 about here

This difference in dispersion represents an overall expansion of the vowel space for hard words. Figure 2 is an F1 by F2 plot of the mean values for each vowel: vowels from "hard" words are plotted using darker slightly larger symbols and vowels from "easy" words are plotted with the lighter symbols. This plot illustrates two characteristics of the data: the overall expansion of the "hard" vowel space, and the tendency for certain vowels to show greater expansion than others.

Insert Figure 2 about here

When the data is analyzed by vowel the differences between vowel types becomes clear: the vowels / i, æ, a, ɔ, u / (point vowels) show the greatest difference between "easy" and "hard words whereas the remainder of the vowels are only slightly expanded or not expanded at all. For some of the vowels / ɪ, ɛ, o,

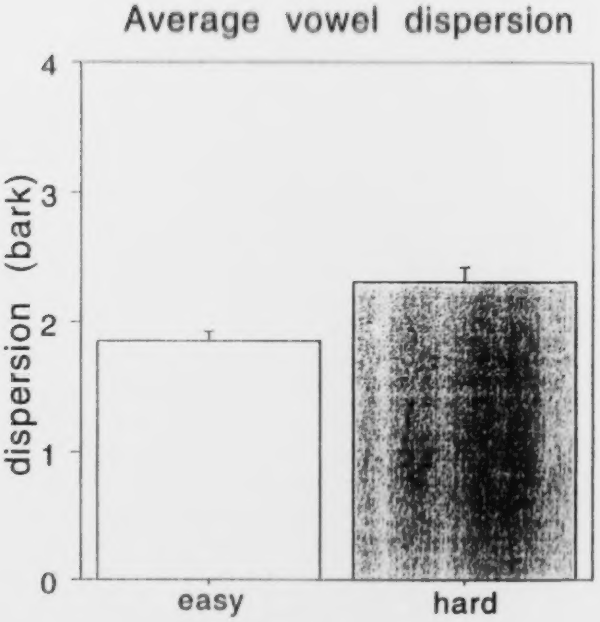


Figure 1. Vowel dispersion for "easy" and "hard" words averaged across talker and vowel, error bars indicate 95% confidence intervals.

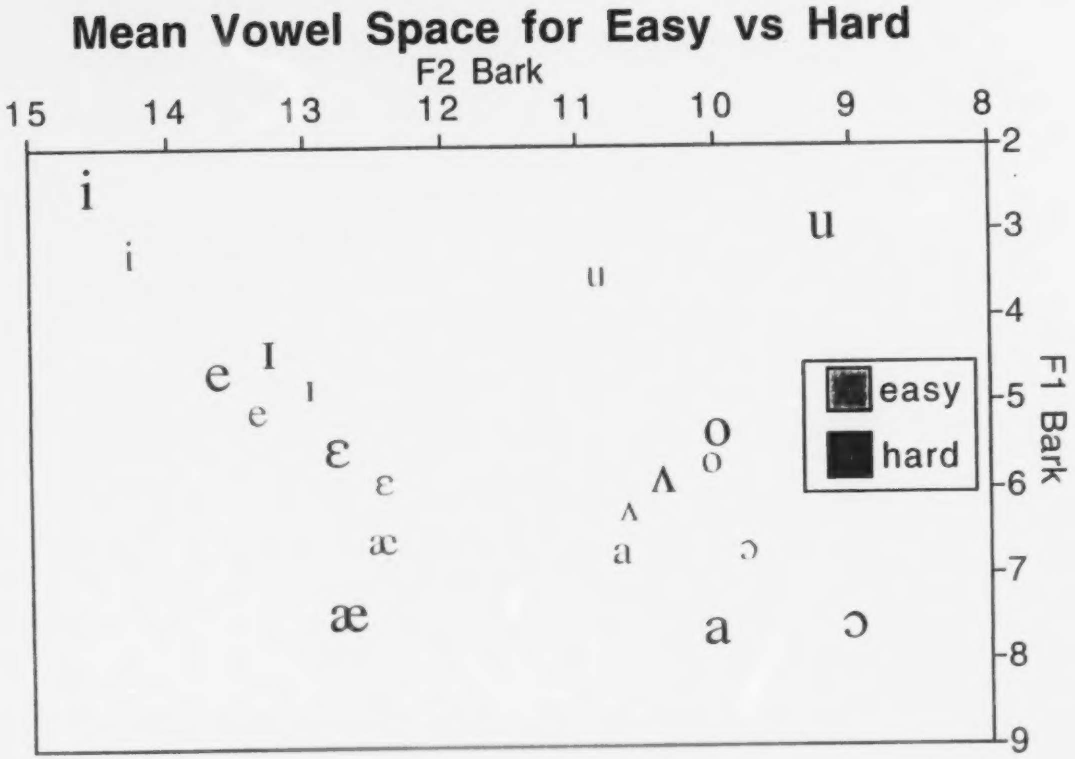


Figure 2. A traditional F1 by F2 vowel plot with F1 on the Vertical axis and F2 on the horizontal axis. Vowel category means are plotted using dark symbols for vowels from “hard” words and lighter symbols for vowels form “easy” words.

ʌ/ there is no reliable difference between the two conditions. Figure 3 illustrates the latter point. It is a plot of dispersion that splits "easy" vs. "hard" by vowel category. There is a marked difference in the height of the dispersion bars between /i, æ, a, ɔ, u/ (point vowels) and the remainder of the vowels.

Insert Figure 3 about here

The difference in dispersion across conditions between the point vowels and the rest of the vowels is in one sense unsurprising. Given the physiological limitations in producing vowels, and given that vowels contrast with each other rather than with the Euclidean center of the space, moving the point vowels while leaving the other vowels fixed maximizes the acoustic distance between vowel of different categories. This stretching of the acoustic space should make the vowel contrasts more salient by increasing the perceptual distance between vowels of differing categories. This finding for English mirrors a crosslinguistic simulation conducted by Liljencrants & Lindblom (1972) that explored the relationship between the number of vowels in a system and the shape of vowel spaces. Their simulation, which took into account the physiological limitations of the vocal tract aimed to find the vowel spaces which maximized the distance between vowels in a two dimensional F1 by F2 space. Their distance formula, given in (2), calculates the repulsive force in a system of vowels using the inverse sum of the squared distances between vowels. With this measure, the closer the different items are the greater the repulsive force.

$$E=\sum_{i=1}^{n-1}\sum_{j=0}^{i-1} 1/r^2_{ij} \qquad (2)$$

$$\text{where } r^2_{ij} \text{ is } \sqrt{(F1_i-F1_j)^2+(F2_i-F2_j)^2}$$

This measure of distance was applied to the set of monophthongs and resulted in a clear difference in repulsive force between the vowel systems from "easy" and "hard" words. Across talkers the "easy" system showed dramatically more repulsive force than the "hard" system, indicating a more expanded vowel space for "hard" words. In looking at individual talkers, this measure reveals striking individual differences. Figure 4 plots repulsive force on the vertical axis by talker (greater repulsive force indicates less overall distance between vowels).

Insert Figure 4 about here

Although the "hard" vowel space showed less repulsive force for all the individuals, some of the individuals exhibited a much greater magnitude of difference. For example, the magnitude of talker M9's difference is several times that of talker F1.

Conclusion

The data supports the hypothesis that vowels from "hard" words are more hyper-articulated than vowels from "easy" words. The expansion of the vowel space occurs in such a way that overall distances between vowels are maximized; only the point vowels, which can move without obscuring the vowel

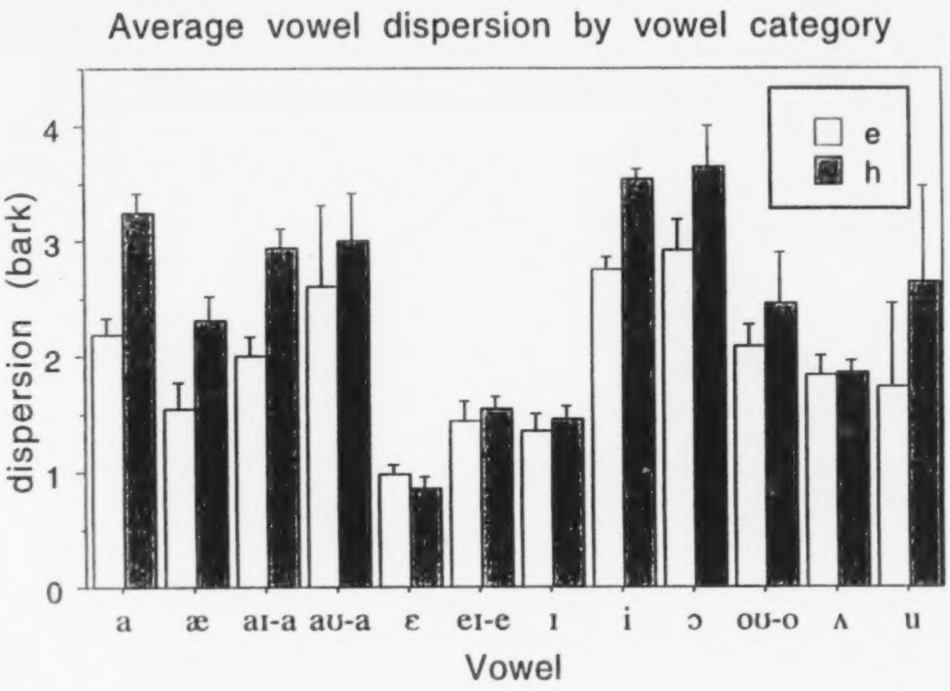


Figure 3. Vowel dispersion for "easy" and "hard" words by vowel type averaged across talker, error bars indicate 95% confidence intervals.

BEST COPY AVAILABLE

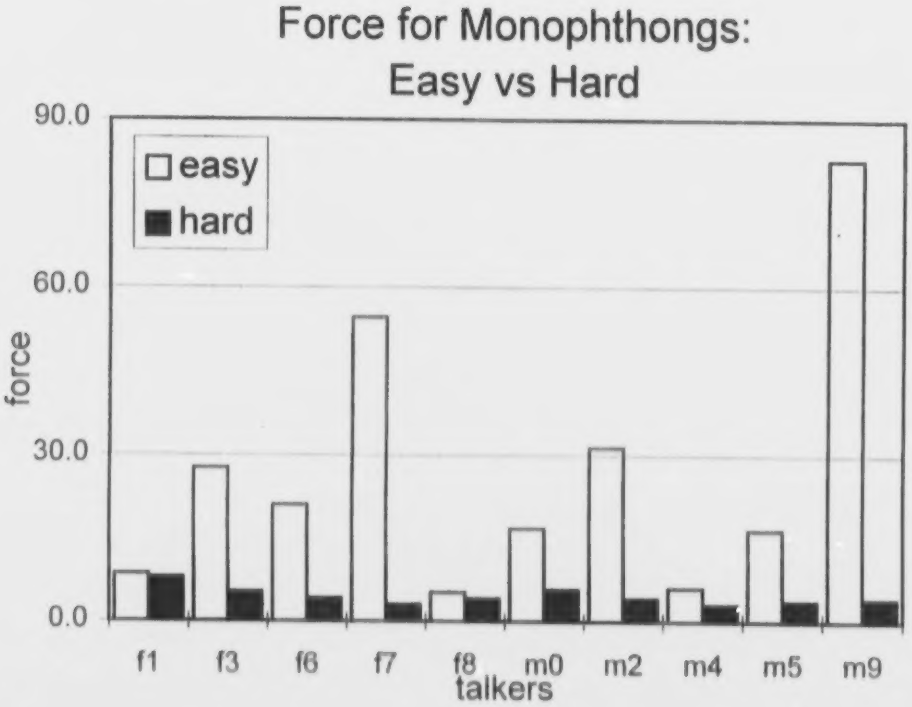


Figure 4. Repulsive force in vowel systems from "easy" and "hard words. Greater force indicates an overall more compact vowel system with less distance between vowels tokens from different categories.

contrasts, become more dispersed while the others remain relatively unchanged across conditions. This finding replicates previous studies' findings that talkers adjust the degree of hyper-articulation to compensate for factors that may impede the intelligibility of a message. This study is novel in that it finds compensatory hyper-articulation for lexical characteristics of individual words. This finding has implications for speech recognition research because it represents a potentially significant reduction in the amount of *random* variability which must be dealt with in an ad-hoc fashion. It also has implications for linguistics, sociolinguistics and historical linguistics in that processes that refer to reduction should take into account lexical properties of words.

It is expected that in follow-up studies, lexical factors will be shown to interact with other processes that promote reduction or hyper-articulation such as information redundancy, noise, the new vs. given status of an utterance, and the familiarity of a talker with a particular hearer. It should be the case that "hard" words will be proportionally more hyper-articulated in conditions where the hearer is expected to have greater difficulty recovering the utterance, and that "easy" words will be proportionally more reduced in conditions where the hearer is expected to have greater ease in recovering the utterance. The individual differences seen in this study are interesting because they indicate non-uniform behavior in the face of relatively uniform differences in factors that affect intelligibility. The individual differences might be thought to be due to differences in the individual's lexicon. Although this is a possibility, it seems unlikely considering the high familiarity of the items used. It may rather indicate differing sensitivity to the effects of lexical factors on intelligibility, different levels of willingness to take advantage of lexical factors, or differential responses to the demands of the recording setup. Further studies of individual differences in hyper-articulatory strategies and their correlations with intelligibility are needed to answer these questions.

References

Anderson, A. H., Bard, E. G., Sotillo, C., Newlands, A., & Doherty-Sneddon, G. (1997). Limited visual control of the intelligibility of speech in face-to-face dialogue. *Perception & Psychophysics*, **39**(4), 580-592.

Balota, D. A., Boland, J. E., & Shields, L. W. (1989). Priming in pronunciation: Beyond pattern recognition and onset latency. *Journal of Memory and Language*, **28**, 14-36.

Bard, E. G., & Anderson, A. H. (1983). The unintelligibility of speech to children. *Journal of Child Language*, **10**, 265-292.

Bolinger, D. (1963). Length, vowel, juncture. *Linguistics*, **1**, 5-29.

Bond, Z. S., & Moore, T. J. (1994). A note on the acoustic-phonetic characteristics of inadvertently clear speech. *Speech Communication*, **14** (4), 325-337.

Bradlow, A. R., Torretta, G. M., & Pisoni, D. B. (1996). Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics. *Speech Communication*, **20**, 255-272.

Bybee, J. L. (1994). A view of phonology from a cognitive perspective. *Cognitive Linguistics*, **5**(4), 285-305.

Byrd, D. (1994). Relations of sex and dialect to reduction. *Speech Communication*, **15**, 39-54.

Chafe, W. (1974). Language and consciousness. *Language*, **50**, 111-133.

Fowler, C. A., & Housum, J. (1987). Talkers' signaling of "new" and "old" words in speech and listeners' perception and use of the distinction. *Memory and Language*, **26**, 489-504.

Goldinger, S. D., & Summers, W. V. (1989). Lexical neighborhoods in speech production: A first report, *Research on Speech Perception Progress Report No. 15* (pp. 331-342). Bloomington, IN: Speech Research Laboratory, Indiana University.

Greenberg, J. H., & Jenkins, J. J. (1964). Studies in the psychological correlates of the sound system of American English. *Word*, **20**, 157-177.

Hawkins, S., & Warren, P. (1994). Phonetic influences on the intelligibility of conversational speech. *Journal of Phonetics*, **22**, 493-511.

Hunnicutt, S. (1985). Intelligibility vs. redundancy—conditions of dependency. *Language and Speech*, **28**, 47-56.

Lane, H., & Tranel, B. (1971). The Lombard sign and the role of hearing in speech. *Journal of Speech and Hearing Research*, **14**, 677-709.

Lane, H., Tranel, B., & Sisson, C. (1970). Regulation of voice communication by sensory dynamics. *Journal of the Acoustical Society of America*, **47**, 618-624.

Lieberman, P. (1963). Some effects of semantic and grammatical context on the production and perception of speech. *Language and Speech*, **6**, 172-187.

Liljencrants, J., & Lindblom, B. (1972). Numerical simulation of vowel quality systems: The role of perceptual contrast. *Language*, **48**, 839-862.

Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H and H theory. In W. Hardcastle & A. Marchal (Eds.), *Speech production and speech modelling*, pp. 403-439. Dordrecht: Kluwer.

Lindblom, B., & Moon, S. J. (1994). Interaction between duration, context and speaking style in English stressed vowels. *Journal of the Acoustical Society of America*, **96**, 40-55.

Lombard, E. (1911). Le signe de l'élévation de la voix. *Annales des maladies de l'oreille, du larynx, du nez et du pharynx*, **37**, 101-119.

Luce, P. A. (1986). *Neighborhoods of words in the mental lexicon*. Unpublished Ph.D. dissertation, Indiana University.

Nusbaum, H. C., Pisoni, D. B., & Davis, C. (1984). Sizing up the Hoosier mental lexicon, *Research on Speech Perception Progress Report No. 10*, . Bloomington, IN: Speech Research Laboratory, Indiana University.

Picheny, M. A., Durlach, N. I., & Braida, L. D. (1985). Speaking clearly for the hard of hearing I: Intelligibility differences between clear and conversational speech. *Journal of Speech and Hearing Research*, **28**, 96-103.

Picheny, M. A., Durlach, N. I., & Braida, L. D. (1986). Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech. *Journal of Speech and Hearing Research*, **29**, 434-446.

Pisoni, D. B., Nusbaum, H. C., Luce, P. A., & Slowiaczek, L. M. (1985). Speech perception, word recognition, and the structure of the lexicon. *Speech Communication*, **4**, 75-95.

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, **26**, 212-215.

Torretta, G. M. (1995). The "easy-hard" word multi-talker speech database: An initial report, *Research on Spoken Language Processing Progress Report No. 20* (pp. 321-333). Bloomington, IN: Speech Research Laboratory, Indiana University.

Uchanski, R. M., Choi, S., Braida, L. D., Reed, C. M., & Durlach, N. I. (1996). Speaking clearly for the hard of hearing IV: Further studies of the role of speaking rate. *Journal of Speech and Hearing Research*, **39**, 494-509.

Zipf, G. K. (1935). *The Psycho-biology of Language*. Boston: Houghton Mifflin.

Zwicker, E., & Terhardt, E. (1980). Analytical expressions for critical-band rate and critical bandwidth as a function of frequency. *Journal of the Acoustical Society of America*, **68**, 1523-1524.

Appendix 1: Words Used in Study

"easy"	"hard"
job	wad
watch	knob
shop	cod
gas	pat
jack	hack
path	hash
five	rhyme
wife	white
vice	lice
mouth	rout
gave	fade
faith	dame
shape	mace
page	sane
chain	wade
death	den
check	wed
leg	pet
peace	bead
deep	teat
teeth	weed
give	kit
thing	hick
ship	kin
thick	mitt
wash	cot
both	goat
vote	moat
food	hoot
young	hum
love	pup
judge	mum
hung	bum
rough	bug

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 21 (1996-1997)
Indiana University

Training Japanese Listeners to Identify English /r/and /l/: Long-term Retention of Learning in Perception and Production[1]

Ann R. Bradlow,[2] Reiko Akahane-Yamada,[3] David B. Pisoni and Yoh'ichi Tohkura[3]

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405

---

[1] This work was supported by NIH-NIDCD Training Grant DC-00012 and NIH-NIDCD Research Grant DC-00111 to Indiana University. We are grateful to Luis Hernandez and Takahiro Adachi for technical support, and to Rieko Kubo and Melissa Kluck for subject running.

[2] Department of Communication Sciences and Disorders, Northwestern University, Evanston, IL

[3] ATR Human Information Processing Research Laboratories, Kyoto, Japan

Training Japanese Listeners to Identify English /r/and /l/: Long-term Retention of Learning in Perception and Production

Abstract. Previous work from our laboratories has shown that monolingual Japanese adults who were subjected to intensive high-variability perceptual training improved in both perception and production of English /r/-/l/ minimal-pairs. This study extended those findings by investigating the long-term retention of learning in both perception and production of this difficult non-native contrast. Results showed that three months after completion of the perceptual training procedure, the Japanese trainees maintained their improved levels of performance on the perceptual identification task. Furthermore, perceptual evaluations by native American English listeners of the Japanese trainees' pretest, post-test, and three-month follow-up speech productions showed that the trainees retained their long-term improvements in the general quality, identifiability, and overall intelligibility of their English /r/-/l/ word productions. Taken together, the results provide further support for the efficacy of high-variability laboratory speech sound training procedures, and suggest an optimistic outlook for the application of such procedures for a wide range of "special populations."

Introduction

Over the past decade, several important advances have been made towards establishing effective laboratory training procedures for modifying the identification of difficult non-native phonetic categories (for recent reviews see Akahane-Yamada, 1996; Jamieson, 1995; Logan and Pruitt, 1995; Pisoni and Lively, 1995; Pisoni, Lively and Logan, 1995). In addition to the benefits that such speech sound training procedures present for second-language learners, this general research agenda also provides important new information regarding the extent to which the adult phonetic system is plastic, and thus capable of undergoing linguistically meaningful modifications. In our laboratories, we have focused our efforts on the acquisition of the English /r/-/l/ contrast by monolingual Japanese speakers. This contrast was selected as a test case for assessing novel approaches to non-native speech contrast training because of its extreme difficulty for Japanese speakers (Goto, 1971; Miyawaki et al. 1975; Mochizuki, 1981; MacKain et al., 1981, Sheldon and Strange, 1982; Yamada and Tohkura, 1992), and because it had been shown in previous studies to be resistant to modification after discrimination training with synthetic CV stimuli (see Strange and Dittmann, 1984).

Accordingly, a laboratory training procedure that included several novel features was developed in our laboratories (Logan, Lively and Pisoni., 1991). In this training procedure, a minimal-pair identification task was used in order to encourage classification into broad phonetic categories, rather than emphasize discrimination of fine-grained within-category acoustic differences. Furthermore, the trainees were presented with naturally produced tokens of English /r/ and /l/ words with the target segment in a variety of phonetic environments. Finally, all training stimuli were uttered by multiple talkers of General American English. In this manner, the trainees were exposed to the full range of category variability that they could expect to encounter in real-world English /r/ and /l/ exemplars. More importantly, the training task closely matched the demands of the identification task used to assess changes in spoken word recognition performance before and after training. The results of several initial training studies using this "high-variability" perceptual training procedure demonstrated that Japanese trainees could acquire robust /r/ and /l/ phonetic categories that generalized to novel talkers and novel tokens (Logan et al., 1991; Lively et al., 1993; Yamada, 1993). Moreover, these changes were retained for several months after the completion of training (see Lively et al, 1994).

More recently, we also showed that the improved English /r/-/l/ identification that resulted from these perceptual training procedures transferred to improved production of English /r/ and /l/ words (Bradlow et al., 1997). Specifically, using a "playback" design, the Japanese trainees' post-test productions of English /r/ and /l/ words were judged by American English listeners to be "better pronounced" than the corresponding pretest productions. The post-test productions were also more accurately identified in a forced-choice minimal-pair identification task than the pretest productions. Thus, the perceptual changes that resulted from the high-variability training procedure extended beyond the perceptual domain and also produced changes in speech production and motor control used in the articulation of these non-native phonetic categories. This finding has provided important new information regarding the relationship between speech perception and production, and suggested that the perceptually-oriented training program resulted in modifications of an underlying perceptuomotor, phonetic representation that is common or shared by both speech perception and production mechanisms.

The goal of the present study was to extend this latest finding by investigating whether the observed changes in speech production were retained for several months after the perceptual identification training was completed. Lively et al. (1994) showed that improvement in perceptual identification was retained for three months after training. We were therefore interested in comparing the retention of production improvement with that of perception improvement. A second goal of this study was to investigate whether the production improvement was also present in an open-set transcription task where the American English listeners were given no clues regarding the identity of the intended word. This perceptual assessment of the production improvement after perceptual identification training would allow us to assess the extent to which the Japanese trainees showed an improvement in overall word intelligibility, in addition to the improvement in general quality and minimal-pair identifiability that we observed in our earlier study (Bradlow et al., 1997). The results of these investigations would allow us to develop a more detailed understanding of the long-term phonetic changes that resulted from the high-variability perceptual identification training procedure.

Methods

Perception Training

The stimuli and procedure used to train these subjects have been described in detail in our earlier papers (see Logan et al, 1991; Lively et al, 1993; 1994; Yamada, 1993; Bradlow et al., 1997). Therefore, in the present report we provide only a brief description of our training methodology, and refer the reader to the previous papers for additional details. The speech stimuli were selected from a large digital database of naturally produced /r/-/l/ minimal-pairs that was originally recorded in the Speech Research Laboratory at Indiana University (see Logan et al., 1991). The pretest stimuli consisted of 16 English minimal-pairs that contrast /r/ and /l/ in four phonetic environments, plus four additional minimal-pairs that contrast other English phonemes (Strange and Dittmann, 1984). These words were all spoken by a male speaker of General American English. The stimuli for the training phase consisted of 68 minimal-pairs that contrast /r/ and /l/ in five phonetic environments. These utterances were spoken by five speakers of General American English (three males and two females). At the post-test phase, subjects were presented with three sets of stimuli: the original pretest stimuli, plus two sets of generalization stimuli. The stimuli for the first test of generalization (TG-1) consisted of an additional 96 words that placed /r/ or /l/ in five different phonetic environments spoken by a new talker (i.e., not one of the talkers that produced the training stimuli). The stimuli for the second test of generalization (TG-2) consisted of an additional 99 words (five phonetic environments) spoken by an old talker (i.e., one of the talkers that produced the training stimuli). In order to assess retention of improved perceptual identification abilities, a three-month follow-up test was administered in which the subjects were tested using the original pretest stimuli as well as the stimuli for the two tests of generalization.

All perception training and testing was done at ATR Human Information Processing Research Laboratories in Kyoto, Japan using individual subject cubicles that were equipped with NeXT workstations and headphones (STAX-SR-Lambda Signature). On each trial, the two members of an English /r/-/l/ minimal-pair appeared on the screen in standard English orthography. The spoken test word was then presented over headphones, and the subjects had ten seconds to identify the stimulus by pressing "1" for the word on the left of the screen or "2" for the word on the right of the screen. During training, feedback was provided in the form of a buzzer (incorrect response) or a chime (correct response). As an additional motivation to perform well on the training task, the trainees received a one yen bonus for each correct response. There was no feedback for the pretest, post-test, tests of generalization, or the three-month follow-up perceptual tests. The training phase took place over a period of 3-4 weeks, during which time the trainees returned to the laboratory 15 times for training sessions.

Speech Production Recordings

At the time of the pretest, post-test, and three-month follow-up phases, the Japanese trainees were also asked to produce a set of 55 English /r/-/l/ minimal-pairs. This set of stimuli included words that placed the target /r/ and /l/ segments in a variety of phonetic environments (see Bradlow et al., 1997 for additional details). Additionally, a set of 25 non-words were recorded. These non-words placed the target /r/ or /l/ adjacent to five vowels (/i, e, a, o, u/) and in a variety of syllable contexts (CV, CCV, VCV, VC, VCC). These non-words were collected primarily for future acoustic analysis, and were therefore omitted from the perceptual evaluation tests by American English listeners. The audio recordings were made in an anechoic chamber at ATR Human Information Processing Research Laboratories. The procedure used to elicit these utterances was an imitation task that presented the subjects with both visual and auditory prompts. For each word, the visual prompt was simply the target English word displayed in standard English orthography on a cardboard panel, and the auditory prompt was a recording of a male speaker of General American English producing the target word. This auditory prompt was provided in order to ensure consistent pronunciation of the rest of the words (aside from the /r/ or /l/) across subjects. Once collected and stored digitally at ATR, these digital speech files were transferred to the Speech Research Laboratory at Indiana University where they were presented to native speakers of General American English for perceptual evaluation.

Perceptual Evaluations of Trainee Productions

Three independent perceptual evaluation tests were carried out in which native speakers of General American English were asked to judge the Japanese trainees' pretest, post-test and three-month follow-up utterances. These playback tests included a preference rating task, a minimal-pair identification task, and an open-set transcription task. For each test, independent groups of ten listeners evaluated the productions of each Japanese subject. Thus, each American English listener evaluated the utterances of only one Japanese subject. Furthermore, no listener participated in more than one evaluation test. All of the perceptual evaluation tests were carried out in the Speech Research Laboratory at Indiana University in Bloomington.

Preference Rating Task.

In the preference rating task, the American English listeners were asked to directly compare the relative phonetic qualities of two versions of a single Japanese subject's productions (e.g. pretest versus post-test, pretest versus three- month). In this task, the listener heard two tokens of an English /r/ or /l/ word (e.g. pretest and post-test), and then indicated on a seven-point scale, which version sounded "better." The target word appeared in standard English orthography on a CRT monitor so that the listener was aware of the Japanese subject's intended pronunciation. A response of "1" indicated that the first version sounded "much better" than the second, a response of "7" indicated that the second version sounded "much better" than the first, and a response of "4" indicated that there was no noticeable difference. The order of presentation of the two versions was counter-balanced across trials. This test provided a very sensitive measure of any improvement in the general quality of English /r/ and /l/ words produced by the Japanese subjects.

Minimal-Pair Identification Task

In the minimal-pair identification task, the American English listeners were asked to identify and categorize the Japanese subjects' productions using a two-alternative forced-choice presentation format. On each trial, the listeners saw the two members of the minimal-pair in standard English orthography on a CRT monitor. They then heard one of the members of the pair spoken by the Japanese subject and responded by identifying the stimulus with one of the two written words. The order of the response alternatives was counter-balanced across trials so that on half the trials the /l/ word was on the left of the CRT monitor, and on the other half, it was on the right of the monitor. This perceptual test provided a quantitative measure of segment-specific improvement in /r/ and /l/ articulation.

Open-Set Transcription Task

The open-set transcription task was a dictation task in which the listener was given no context regarding the identity of the Japanese subject's utterance. In this test, the listeners heard a word spoken by the Japanese subject and then responded by typing what they heard into the keyboard. The responses were scored such that a word was counted as correctly transcribed if, and only if, the transcription exactly matched the intended word (aside from any obvious typographical or spelling errors). This test provided a strict test of overall word intelligibility without context.

Taken together, the three perceptual assessment tests provided us with a converging set of behavioral measures of the changes in speech production that resulted from the perceptual identification training procedure. These measures allowed us to assess the extent of the changes in general quality (preference rating task), in segment-specific articulation (minimal-pair identification task), and overall speech intelligibility (open-set transcription task).

Subjects

A group of eleven native Japanese speakers (five females and six males) served as the trained subjects. Of these eleven trained subjects, nine returned for the three-month follow-up test. They ranged in age from 19 to 22 years and were recruited from Doshisha University in Kyoto, Japan. None had ever lived abroad or had any special English language training. A comparable group of twelve subjects (six females and six males) served as untrained controls. Of these twelve Japanese controls, seven returned to the laboratory for the three-month follow-up test. Finally, all of the native American English listeners who served as judges of the Japanese subjects' pre- and post-test utterances were recruited from the Indiana University student population. For each Japanese subject, separate panels of ten American English listeners served as judges in each of the production evaluation tests.

Results

Perceptual Learning

Figure 1 shows the perceptual identification accuracy scores at pretest, post-test, and at the three-month follow-up test for the nine trained (left panel) and seven control (right panel) subjects who participated in all three phases of the study (pretest, post-test, and three-month follow-up). Table I provides the complete set of identification scores for all of the individual trained and control subjects on the pretest, post-test, three-month follow-up test, as well as on the two tests of generalization that were administered only at the post-test and three-month follow-up phases (see also Bradlow et al., 1997). As shown in this figure, the trained subjects improved substantially above pretest scores in their ability to identify English /r/ and /l/ words at the post-test and three-month follow-up phases, whereas the control subjects showed no change in perceptual identification accuracy across these conditions. A two-factor repeated-measures ANOVA, with test (pre, post, 3-month) as the repeated measure and group (trained, control) as the between-groups factor showed a main effect of test ($F(2,28)=13.851$, $p<.001$), and a main effect of group ($F(1,14)=5.65$, $p=.032$). The group x test interaction was also significant ($F(2,28)=15.17$, $p<.001$) due to the difference in accuracy scores across tests for the trained group, but not for the control group. Paired t-tests showed a significant improvement for the trained group from pretest to post-test ($t(8)=-7.392$, $p<.005$), and from pretest to three-month follow-up ($t(8)=-3.905$, $p<.005$). There was no difference in performance for the trained group between the post-test and three-month follow-up scores. Furthermore, there was no difference between the trained and control groups' pretest accuracy scores, indicating that the two groups were indeed comparable at the time of pretest.

Insert Figure 1 about here.

An examination of the individual subject data (see Table I) shows that, of the nine trained subjects who returned to the laboratory for the three-month follow-up test, seven maintained a level of performance that was at least eight percentage points higher than their pretest level of performance. Only two subjects showed a decrease in identification accuracy back to their pretest level of performance. Furthermore, at both the post-test and three-month follow-up phases, the gains made in perceptual identification accuracy scores generalized to novel items and novel talkers (see Table I). Thus, the information these subjects learned in the training generalized well beyond the specific items used in the perceptual training task.

These data replicate the earlier findings of Lively et al. (1993) who showed that the group of trained subjects maintained the improved level of identification ability even three months after perceptual identification training was completed. In contrast, the group of control subjects showed no change in perceptual identification accuracy from pretest to post-test or to the three-month follow-up test. Having established that the perceptual training procedure produced long-term changes in these Japanese trainees' ability to identify English /r/ and /l/ words, we now turn to an examination of the long-term changes in production of English /r/ and /l/ words that resulted from the perceptual identification training.

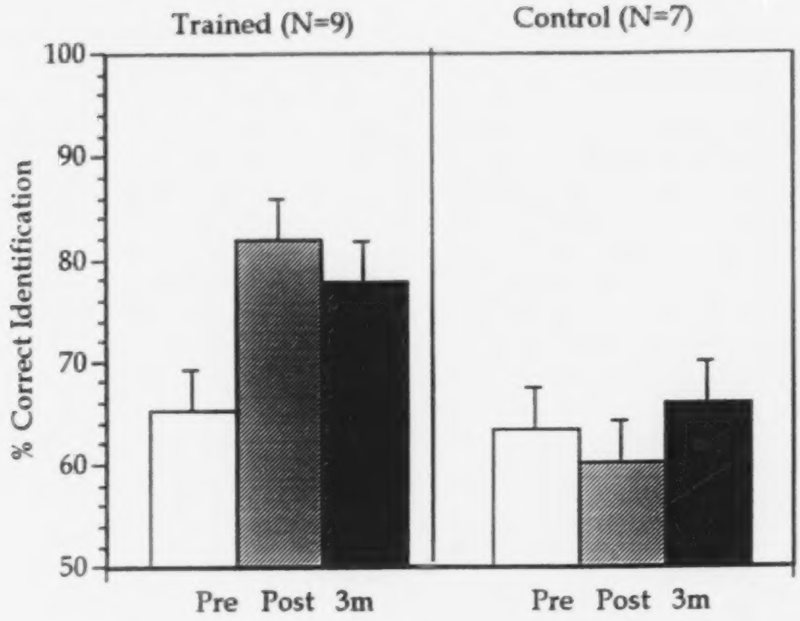


Figure 1. Percent correct perceptual identification performance for trained (left panel) and control (right panel) subjects at pretest, post-test, and three month follow-up for the subjects who participated in all three test phases. The error bars represent one standard error from the mean.

Table I.

Pretest, post-test, and three-month follow-up perceptual identification accuracy scores for all Japanese trained and control subjects. Also shown are the scores for the two tests of generalization (TG1=new words, new talker; TG2=new words, old talker) that were given at the post-test and three-month follow-up phases.

Trained	pretest	post-test	post-TG1	post-TG2	3-month	3m-TG1	3m-TG2
1	67.19	81.25	89.90	85.42	82.10	80.81	85.42
2	85.94	95.31	96.97	97.92	95.31	95.96	96.88
3	56.25	78.12	59.60	50.00	57.81	48.49	51.04
4	82.81	96.88	96.97	96.88	90.63	95.96	96.88
5	65.63	76.56	78.79	86.46	67.19	68.67	70.83
6	56.25	76.56	81.82	72.92	75.00	79.80	68.75
7	51.56	59.38	66.67	61.46	59.38	59.60	56.25
8	68.75	92.19	95.96	89.58	85.94	89.90	84.38
9	56.25	62.50	62.63	61.46	—		
10	57.81	84.38	88.89	89.58	89.06	82.82	78.13
11	67.19	92.19	93.94	87.50	—		
mean	65.06	81.39	82.92	79.92	78.05	78.00	76.51
Control	**pretest**	**post-test**	**post-TG1**	**post-TG2**	**3-month**	**3m-TG1**	**3m-TG2**
1	57.81	59.38	54.55	57.29	—	—	—
2	64.06	60.94	57.58	59.38	65.63	61.62	55.21
3	62.50	51.57	58.59	58.33	—	—	—
4	67.19	62.50	68.69	65.63	67.19	67.68	72.92
5	62.50	53.13	59.60	53.13	—	—	—
6	73.44	62.50	66.67	64.58	—	—	—
7	71.88	71.88	69.70	76.04	68.75	62.63	61.46
8	54.69	48.44	48.48	52.08	54.69	55.56	57.29
9	57.81	53.13	64.65	54.17	65.63	61.62	62.50
10	54.69	56.25	49.50	56.25	—	—	—
11	67.19	62.50	57.58	66.67	70.31	56.57	63.54
12	73.44	68.75	61.62	68.75	70.31	61.62	70.83
mean	63.93	59.25	64.65	57.29	66.07	61.04	63.39

Production Improvement

Figure 2 shows the results of the preference rating task in which American English listeners directly compared the Japanese subjects' pretest versus post-test utterances (panels (a) and (c)) and the pretest versus the three-month follow-up utterances (panels (b) and (d)). The data shown here are for the nine trained subjects (upper panels) and the seven controls (lower panels) that participated in all three phases of the study. Recall that in this preference rating test the American English listeners heard two tokens (e.g. one pretest token and one post-test token) of a given word spoken by one Japanese subject, and they responded by indicating on a seven-point scale which version was "better articulated." In order to take into account the counter-balanced presentation order of the two versions, the data were all recoded so that a response of 1, 2, or 3 always indicated a preference for the pretest token and a response of 5, 6, or 7 always indicated a preference for the post-test (or three-month follow-up) token. The figure shows the distribution of rating responses across the seven response categories represented as percentages of the total number of responses from all listeners.

Insert Figure 2 about here.

As shown in Figure 2, the distribution of the ratings for the Japanese trained subjects' utterances was skewed in favor of consistently higher ratings for the test that compared the pretest versus the post-test utterances (panel (a)), as well as for the test that compared the pretest versus the three-month follow-up utterances (panel (b)). This tendency towards higher ratings indicated that the American English listeners reliably preferred the Japanese trainees' post-test and three-month follow-up utterances over their pretest utterances. In contrast, for both tests with the control subjects utterances (panels (c) and (d)), the ratings were normally distributed across the seven response categories, indicating no general preference for either the post-test or the three-month follow-up utterances over the pretest utterances. From this pattern of ratings, we conclude that the tokens from the control subjects were indiscriminable to native speakers of English.

Chi-square statistics were performed on the distribution of responses across all seven response categories using the distribution of responses to the control subjects' productions as the expected distributions, and the distribution of responses to the trained subjects productions as the observed distributions. These analyses yielded highly significant chi-squares for the pretest versus post-test comparison (chi-square=3033.46, $p(6)<.001$), and for the pretest versus three-month follow-up comparison (chi-square=1938.87, $p(6)<.001$). Taken together, these data provide evidence that even three months after training was completed, the perceptual learning was retained and the improved general quality of the Japanese trainees' /r/ and /l/ words was still intact.

The second production evaluation test, the minimal-pair identification test, allowed us to investigate changes in production that were specific to /r/ and /l/ articulation. In this test, American English listeners identified and categorized each word produced by a Japanese subject as either the /r/ or the /l/ member of an /r/-/l/ minimal-pair. For each Japanese trainee, two separate tests were carried out using separate groups of American English listeners. In the first, the pretest and post-test productions were presented; in the second, the pretest and three-month follow-up productions were presented. Thus, the post-test and three-month follow-up productions were each identified in data collection sessions that also included the pretest productions. A comparison of the identification accuracies for the two presentations of the pretest productions show no significant difference between the session with the post-test productions and the session with the three-month follow-up productions, therefore in the final data analysis the two set of pretest scores were averaged.

Table II shows the identification accuracy scores from the American English listeners' judgments of the pretest, post-test, and three-month follow-up recordings from the nine Japanese trainees who returned three month after training was completed. A one-factor repeated measures ANOVA with test (pretest, post-test, three-month) as the repeated measure, showed a main effect of test ($F(2,16) = 6.381$, $p<.01$). Paired t-tests showed a significant increase in identification accuracy between pretest and post-test ($t(8)=-2.516$, $p<.05$), and between pretest and three-month follow-up test ($t(8)=-2.601$, $p<.05$), but no difference between post-test and three-month follow-up. These data demonstrate that, even three-months after the perceptual identification training was completed, the segment-specific improvement in /r/ and /l/ articulation was retained by these subjects.

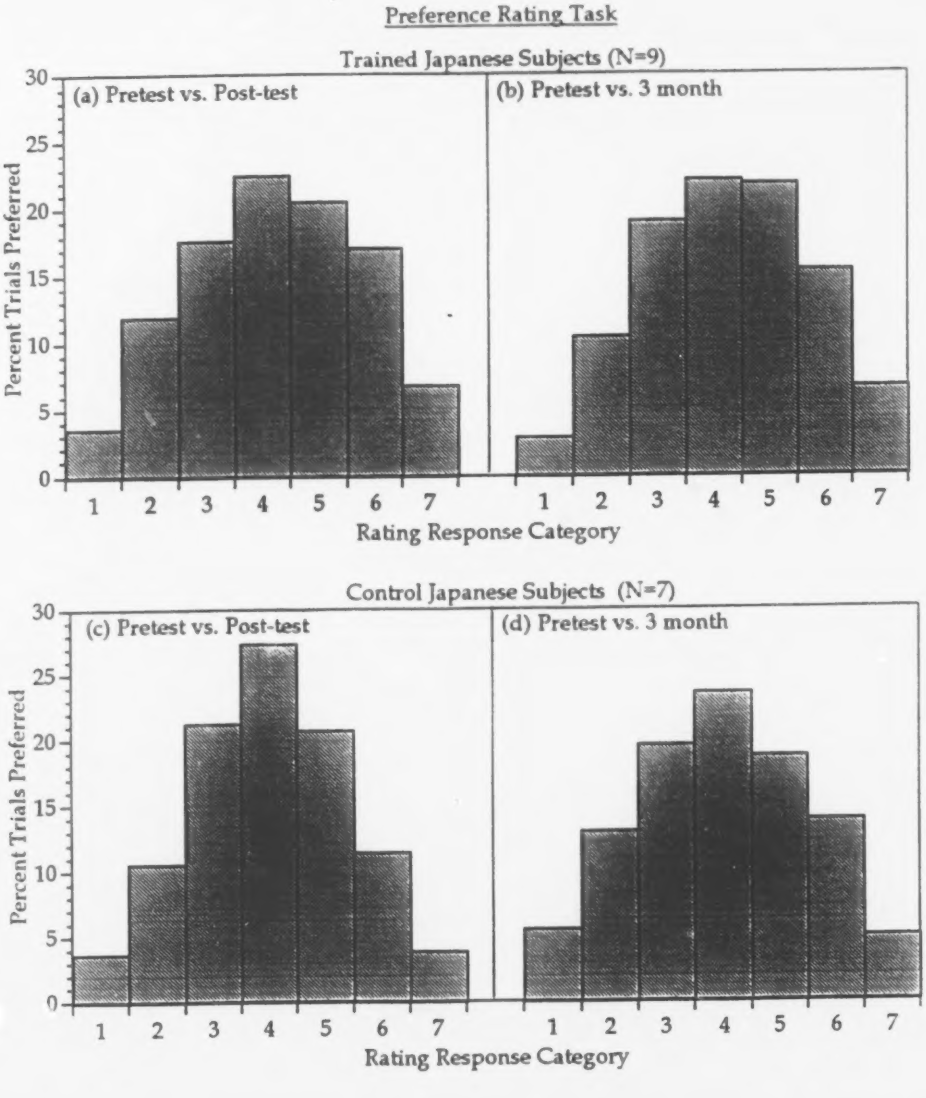


Figure 2. Distribution of responses across the seven response categories for the preference rating tests comparing the trained (top) and control (bottom) subjects' pretest versus post-test productions on the right (panels (a) and (c), respectively), and pretest versus three-month follow-up productions on the left (panels (b) and (d), respectively). A response of "1" indicated that the pretest version was preferred over the post-test or three-month token, "4" indicated no noticeable difference between the two tokens, and "7" indicated that the post-test or three-month follow-up token was preferred over the pretest token. Data are shown for only those subjects who participated in all three phases of the study, pretest, post-test and three-month follow-up.

546

Table II.

Individual trainee minimal-pair identification and open-set transcription scores as judged by American English listeners at pretest, post-test and three-month follow-up.

	Minimal-pair Identification			Open-set Transcription		
Trainee	pretest	post-test	3-month	pretest	post-test	3-month
1	55.95	73.00	80.05	26.01	35.97	39.18
2	95.75	95.18	97.23	53.87	55.28	57.00
3	60.43	65.41	73.86	27.03	36.63	38.27
4	98.50	98.95	98.32	74.09	71.18	83.73
5	59.86	60.91	59.23	34.39	36.10	36.18
6	62.50	72.14	73.55	36.99	42.27	45.27
7	60.75	60.18	58.05	30.93	27.94	29.27
8	75.73	81.32	85.82	47.10	54.10	59.00
9	60.00	76.09	—	34.64	38.29	—
10	56.64	62.55	68.27	29.26	34.38	35.18
11	56.50	60.55	—	30.79	28.20	—
mean	67.51	73.30	77.15	38.65	41.85	47.01

The control subjects' three-month follow-up recordings were not submitted to the minimal-pair identification task (or the open-set transcription production evaluation task) since the data from the preference rating task indicated no discriminable change in the overall quality of the control subjects' productions from pretest to post-test, or from pretest to three-month follow-up test. Furthermore, in an earlier paper (Bradlow et al., 1997) we reported the results of the minimal-pair identification task for the control subjects pretest and post-test productions (n=12), which showed no difference in the American English listeners' identification accuracies for the pretest and post-test productions. Thus, there was strong a priori evidence that the control subjects' productions did not change at all from pretest to post-test to three-month follow-up test. We therefore eliminated their productions from any further production evaluation tests under the assumption that if there were no reliable differences at post-test there would also be no differences at the three-month follow-up test.

The third, and final, production evaluation test allowed us to examine overall word intelligibility of the Japanese trainees' pretest, post-test and three-month follow-up productions in the absence of any context or response constraints. In this open-set transcription task, the American English listeners heard a word spoken by a Japanese trainee, and then typed what they heard into the keyboard. Table II shows the percent correct transcription scores for the trainees' pretest, post-test and three-month follow-up productions. This production evaluation test provides a very stringent measure of overall word intelligibility using an open-set response format. Thus, the overall percent correct transcription scores were considerably lower than the percent correct identification scores that were obtained in the minimal-pair identification test which had a chance level of 50%. Nevertheless, we observed a significant improvement in the overall intelligibility of the Japanese trainees' productions from pretest to post-test, and this improved level of performance was maintained even three months after perceptual identification training was completed.

A one-factor repeated measures ANOVA with test (pretest, post-test, three-month) as the repeated measure, showed a main effect of test ($F(2,16) = 10.576$, $p<.005$). Paired t-tests showed a significant increase in identification accuracy between pretest and post-test ($t(8)=-2.356$, $p<.05$), and between pretest and three-month follow-up test ($t(8)=-4.155$, $p<.05$). We also found a significant difference between the post-test and three-month follow-up test ($t(8)=-2.583$, $p<.05$), such that the three-month follow-up productions were more accurately transcribed than the post-test productions. The reason for this increase is unclear at this time. However, the important finding for our purposes is that both the post-test and three-month follow-up productions were more accurately transcribed than the pretest productions and there was no decrease in the intelligibility scores after three-months.

In summary, the three perceptual evaluation tests provided independent, and converging support for the claim that the "high variability" perceptual identification training procedure produced long-term changes in the Japanese trainees' control over production of English /r/ and /l/ words. The first perceptual test, the preference rating task, was a highly sensitive, relative measure of the general quality of the Japanese subjects' productions. The second perceptual test, the minimal-pair identification task, was a segment-specific probe that provided direct evidence for improvement in the Japanese trainees' /r/ and /l/ articulations. The final test, the open-set transcription test, provided a measure of improvement in overall word intelligibility in the absence of any contextual cues for the listener regarding the identity of the target word. Thus, each of these production evaluation tests provided us with a different assessment of the changes in speech production that the perceptual identification training procedure produced in the Japanese trainees. The improvements were both general and segment-specific, and resulted in higher overall word intelligibility that was retained even three months after training was completed.

Discussion

The primary goal of this study was to investigate and assess the nature of changes in perceptual identification and production of English /r/-/l/ minimal-pairs following intensive perceptual identification training and to measure the retention of this knowledge over time. The findings showed that the "high-variability" perceptual training procedure did indeed produce long-term modifications in both perception and production of a difficult non-native phonetic contrast. These findings demonstrate the importance of stimulus variability for the acquisition and retention of fine phonetic details about these segmental contrasts. Furthermore, the transfer and retention of knowledge across receptive and expressive domains implies a close link between speech perception and production during perceptual learning of novel phonetic contrasts.

At this point, we can identify three major generalizations regarding speech sound learning that have emerged from our efforts to train Japanese speakers to acquire the English /r/-/l/ contrast in a laboratory setting. First, the consistent success of the "high-variability" training procedure demonstrates that the adult phonetic system displays sufficient neuro-plasticity to undergo substantial modification through laboratory listening training alone. However, it is also important to note that the Japanese trainees who participated in our studies have consistently failed to reach native-like abilities to identify English /r/ and /l/. Thus, although the adult phonetic system apparently maintains the ability to change in response to novel stimuli, it also appears to be subjected to certain limitations imposed by the native language phonetic system.

Second, the robust nature of the perceptual learning exhibited by the trainees in our studies has established that the high-variability training approach is an effective means of producing generalized long-term changes in the underlying phonetic system. Specifically, the improvements in /r/-/l/ identification generalized to novel items and novel talkers, and this knowledge was retained for at least three months after training. The key elements of this training approach that are apparently responsible for the robust learning are the stimuli (i.e., a wide range of /r/ and /l/ exemplars produced by multiple talkers) and the task (i.e., a minimal-pair identification task that encourages classification into broad phonetic categories rather than a discrimination task that encourages perception of fine-grained within-category differences).

Third, the learning produced via the perceptual modality produces long-term modifications to both perception and production of the trained contrast, suggesting that changes to the underlying phonetic system occur at a level of representation that is common to both perception and production. In other words, perceptual training alone produces generalized changes that affect diverse speech processing operations, all of which are retained for several months after training is completed.

The overall pattern of results suggests a very encouraging scenario for the design and application of laboratory speech sound training procedures for second-language learners, as well as for other "special populations" who exhibit difficulties with speech sound perception and production. For example, in a recent investigation of the factors that correlate with superior performance in aural-oral language acquisition by prelingually deafened children with cochlear implants, Pisoni et al. (1997) reported a strong positive correlation between performance on a word recognition test and performance on a test of speech intelligibility. Similarly, Stark and Heinz (1996) found that impaired stop consonant perception by language-impaired children relative to normal children was associated with the presence of corresponding speech articulation errors. Furthermore, Yamada et al. (1994) found a positive correlation between perception and production of the English /r/-/l/ contrast in a large group of Japanese speakers. Thus, available data from a variety of populations that exhibit phonological problems (including second-language learners, pediatric cochlear implant users, and language impaired children) suggest that performance on speech perception tasks tends to correlate positively with performance on corresponding speech production tasks. The present investigation extends this general finding by establishing a perception-production link such that successful perceptual learning leads directly to corresponding improvement in speech production, specifically speech motor control and articulation. Taken together, these results support the claim that phonological acquisition via auditory-perceptual input involves concurrent development in both speech perception and speech production. Moreover, our findings suggest that the high-variability training procedure holds great promise as a general approach to the development of laboratory training procedures for the acquisition of difficult phonological categories in a wide range of "phonologically disabled" populations.

References

Akahane-Yamada, R. (1996). Learning non-native speech contrasts: What laboratory training studies tell us. *Proceedings of the Acoustical Society of Japan Fall meeting,* Honolulu, Hawaii.

Bradlow, A. R., Pisoni, D. B., Akahane-Yamada, R., & Tohkura, Y. (1997). Training Japanese listeners to identify English /r/ and /l/: IV. Some effects of perceptual learning on speech production, *Journal of the Acoustical Society of America,* **101**, 2299-2310.

Goto, H. (1971). Auditory perception by normal Japanese adults of the sounds 'l' and 'r'. *Neuropsychologia*, **9**, 317-323.

Jamieson, D. G. (1995). Techniques for training difficult non-native speech contrasts. *Proceedings of the International Congress of Phonetic Sciences,* **4**, 100-107.

Lively, S. E., Logan, J. D., & Pisoni, D. B. (1993). Training Japanese listeners to identify English /r/ and /l/: II. The role of phonetic environment and talker variability in learning new perceptual categories. *Journal of the Acoustical Society of America*, **94**, 1242-1255.

Lively, S. E., Pisoni, D. B., Yamada, R. A., Tohkura, Y., & Yamada, T. (1994). Training Japanese listeners to identify English /r/ and /l/: III. Long-term retention of new phonetic categories. *Journal of the Acoustical Society of America*, **96**, 2076-2087.

Logan, J. D., Lively, S. E., & Pisoni, D. B. (1991). Training Japanese listeners to identify English /r/ and /l/: A first report. *Journal of the Acoustical Society of America*, **89**, 874-886.

Logan, J. S. & Pruitt, J. S.(1995). Methodological issues in training listeners to perceive non-native phonemes. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language speech research*, (pp. 351-378). Timonium, MD: York Press.

Mackain, K. S., Best, C. T., & Strange, W. (1981). Categorical perception of English /r/ and /l/ by Japanese bilinguals. *Applied Psycholinguistics*, **2**, 369-390.

Miyawaki, K., Strange, W., Verbrugge, R., Liberman, A. M., Jenkins, J. J., & Fujimura, O. (1975). An effect of linguistic experience: The discrimination of [r] and [l] by native speakers of Japanese and English. *Perception and Psychophysics*, **18**, 331-340.

Mochizuki, M. (1981). The identification of /r/ and /l/ in natural and synthesized speech. *Journal of Phonetics*, **9**, 283-303.

Pisoni, D. B., Svirsky, M. A., Kirk, K. I., & Miyamoto, R.T. (1997). Looking at the "stars:" A first report on the intercorrelations among measures of speech perception, intelligibility, and language in pediatric cochlear implant users. Paper presented at the Vth International Cochlear Implant Conference, New York City, NY, May 1-3, 1997.

Pisoni, D. B., Lively, S. E., & Logan, J. S. (1994). Perceptual learning of non-native speech contrasts: Implications for theories of speech perception. In H. C. Nusbaum and J. Goodman (Eds.). *Development of speech perception: The transition from speech sounds to spoken words*, (pp. 121-166). MIT Press: Cambridge, MA.

Pisoni, D. B. & Lively, S. E. (1995). Variability and invariance in speech perception: A new look at some old problems in perceptual learning. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language speech research*, (pp. 433-462). Timonium, MD: York Press.

Sheldon, A., & Strange, W. (1982). The acquisition of /r/ and /l/ by Japanese learners of English: Evidence that speech production can precede speech perception. *Applied Psycholinguistics*, **3**, 243-261.

Stark, R.E. & Heinz, J.M. (1996). Perception of stop consonants in children with expressive and receptive-expressive language impairments. *Journal of Speech and Hearing Research*, **39**, 676-686.

Strange, W., & Dittmann, S. (1984). Effects of discrimination training on the perception of /r/-/l/ by Japanese adults earning English. *Perception and Psychophysics*, **36**, 131-145.

Yamada, R. A. (1993). Effects of extended training on /r/ and /l/ identification by native speakers of Japanese. *Journal of the Acoustical Society of America,* 93, Pt. 2, 2391.

Yamada, R. A. & Tohkura, Y. (1992). The effects of experimental variables on the perception of American English /r/ and /l/ by Japanese listeners. *Perception and Psychophysics,* **52**, 376-392.

Yamada, R. A., Strange, W., Magnuson, J. S., Pruitt, J. S., & Clarke, W. D. III (1994). The intelligibility of Japanese speakers' productions of American English /r/, /l/, and /w/, as evaluated by native speakers of American English. In *Proceedings of the International Conference of Spoken Language Processing*. (pp. 2023-2026). Acoustical Society of Japan: Yokohama.

A Preliminary Acoustic Study of Errors in Speech Production[1]

Stefan Frisch and Richard Wright

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405

---

[1] This work supported by NIH-NIDCD Training Grant DC00012 to Indiana University. This paper was presented as a poster at the 133rd meeting of the Acoustical Society of America, June 1997.

A Preliminary Acoustic Study of Errors in Speech Production

Abstract. Phonological speech errors provide important psycholinguistic evidence for the representations of phonological theory. In an electromyographic (EMG) study of experimentally induced phonological speech errors, Mowrey and MacKay (1990) found that speech errors frequently occur at a sub-featural, gestural level, with no apparent effect on the percept of the word. Based on these gradient errors, they argue against speech errors as evidence for the segmental unit. Mowrey and MacKay's study considered the activity of a single muscle, and thus was unable to determine whether single gestures acted independently of gestural constellations, which may be equivalent to traditional segmental units. This study is a preliminary report from an ongoing acoustic analysis of speech errors. The data are tape recordings of an error inducing experiment using nonsense tongue twisters. Recordings of a single speaker producing four different tongue twisters targeting /s/ and /z/, e.g. sit zap zoo sip, were digitized and analyzed. Some errors involved multiple changes in acoustic properties, including simultaneous changes in periodicity, amplitude of frication, and duration, while others involved a subset of these properties. This evidence suggests that errors can occur at both the single gesture level, affecting non-contrastive acoustic properties, and at the level of the gestural complex or segment, creating a perceptible, linguistically contrastive change.

Introduction

This study is an investigation of sub-lexical phonological speech errors using acoustic-phonetic measures. By using acoustic methods, we intended to evaluate evidence for phonological segments based on speech errors, and to uncover new characteristics of the organization of the speech production mechanism which would not be discovered using traditional transcriptional evidence.

Traditionally, speech error data is collected and analyzed using only phonetic transcriptions. Speech errors are collected either opportunistically, in 'natural error corpora,' or experimentally, from speech error inducing procedures such as the SLIPS priming technique (Baars, Motley, & MacKay, 1975) or tongue twisters (e.g., Shattuck-Hufnagel, 1992). Based on traditional data collection, researchers have claimed that most errors occur at the level of the phoneme or feature (Wickelgren, 1965, Fromkin, 1971). In addition, one 'law' of speech errors is that erroneous utterances are phonotactically grammatical (Wells, 1951; Fromkin, 1971).

Mowrey and MacKay (1990) analyzed electromyographic (EMG) recordings of tongue twisters to evaluate these two claims. They found that EMG activity during tongue twisters showed gradient speech errors. There was a range of muscle activation from none to that equivalent to a normal production for an intruding segment. For example, in tongue twisters such as *Bob flew by Bligh bay*, they found gradient activation of the lingual transversus-verticalis complex, used in the lingual gesture of [l], in the productions of *bay*. Muscle activation was found in productions which were perceptually normal, indicating that a gradient error may occur which is auditorily undetectable.

They conclude that gradient errors which would not be detected by traditional error collection techniques often occur (see also Laver, 1979; Boucher, 1994). Errors which occur on a continuum of muscle activation undermine arguments that the majority of speech errors occur at the phonological level of the phoneme or feature. In addition, such errors violate the law that speech errors obey phonotactic grammaticality under any non-trivial interpretation of this generalization.

Mowrey and MacKay (1990) studied single muscle fiber activation, and did not examine the acoustic properties of their anomalous utterances. Thus, it is impossible to determine whether the gradient activation they found occurred in all fibers of the muscles involved in a linguistically significant gesture, or whether some higher level monitoring in the production component utilized agonistic or antagonist fibers to insure an auditorily normal outcome in cases of gradient errors. We propose that an acoustic analysis of speech errors which considers several dimensions upon which a linguistic contrast is based can reveal the full range of variation in the speech error data. Our finding is that errors may occur on a single acoustic dimension as the result of a single gestural error, or errors may have simultaneous changes on several independent dimensions, involving an entire constellation of gestures.

Methods

Data

Recordings from a speech error experiment (Frisch, 1996) were analyzed acoustically and compared to the transcriptions of the experimental session which were used to score productions as errors. The original experiment had 88 tongue twisters involving a variety of target consonants, all of which were onsets of monosyllables. The data we analyzed acoustically consisted of four tongue twisters targeting [s] and [z], each repeated six times. The twisters, in the order they were presented in the experiment, are given in (1).

(1) sit zap zoo sip
sung zone Zeus seem
zit sap sue zip
zig suck sank zilch

We chose tongue twisters involving [s] and [z] as they generated many errors in the experiment and were difficult to transcribe.

We initially conducted a qualitative analysis of 6 participants of the original 21 participants in the experiment (the first six participants). The results we present here are detailed measurements for one representative participant (Participant 2). To date, we have made measurements of two other participants (Participants 1 and 3) and the overall patterns observed are analogous to those for participant two. All measurements were made from waveforms, with accompanying spectrograms for reference. Table 1 shows the general acoustic characteristics of [s] and [z] we analyzed.

Table 1

Acoustic characteristics of [s] and [z]

Characteristic:	[s]	[z]
Duration (Klatt 1976)	Long	Short
Periodicity	Aperiodic	Periodic
Frication amplitude (Strevens 1960, Pickett 1980)	Greater	Lesser
Vowel onset	Sharp	Gradual

Measurements

Four measurements intended to capture the major differences between [s] and [z] were made:

1. DURATION - the duration of the fricative noise, including overlap with the preceding or following vowels.
2. %VOICING - the fraction of the total duration which contained voicing.
3. WINDOW AMPLITUDE - the RMS amplitude of frication noise of a 50ms window surrounding the amplitude peak of the fricative noise. The signal was high-pass filtered at 2kHz to remove the energy contributed by the periodic signal.
4. VOWEL RISE TIME - the time from the end of the fricative to the first vowel amplitude plateau.

These measurements are demonstrated for [s] and [z] in Figure 1.

Insert Figure 1 about here

Results

Gross Characteristics of the Data

Overall, the four measurements differentiate [s] and [z] for Participant 2. Means for Participant 2's productions for all measurements, with error bars showing standard error, are given in Figure 2.

Insert Figure 2 about here

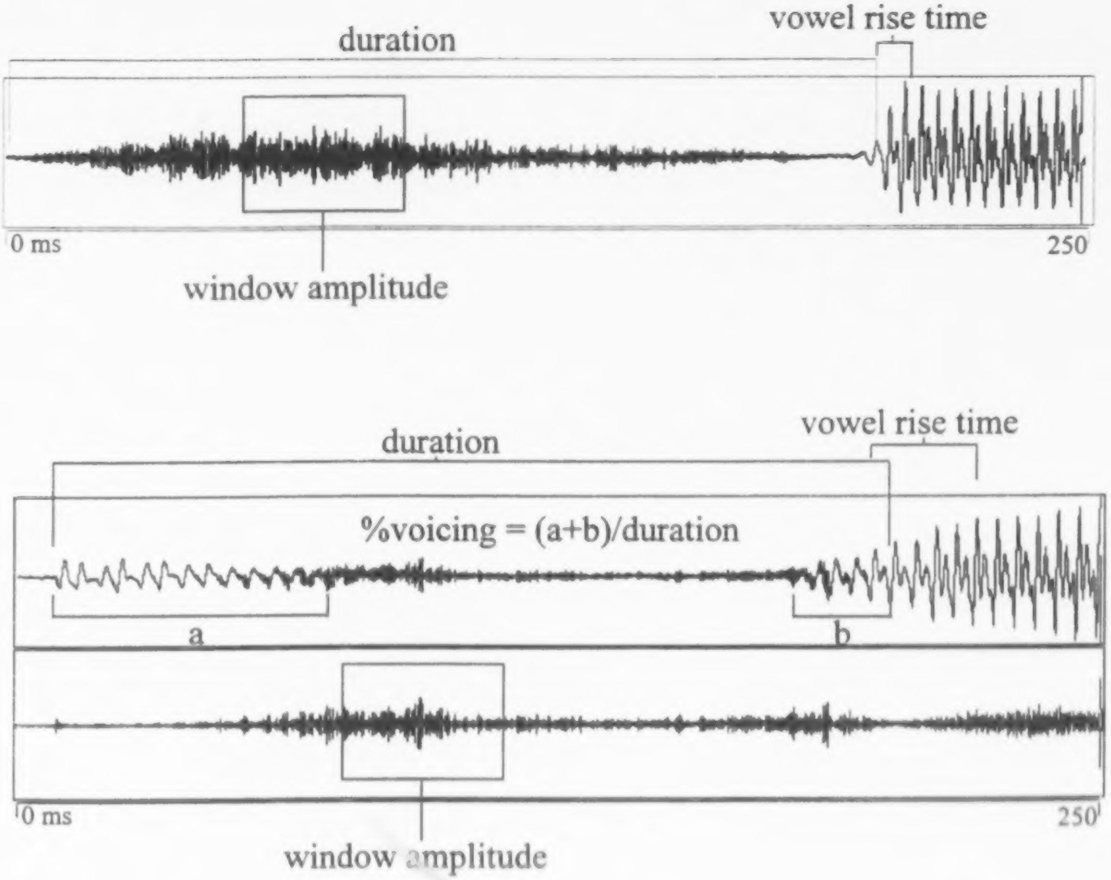


Figure 1. Sample measurements of [s] in *sip* (top) and [z] in *zilch* (bottom).

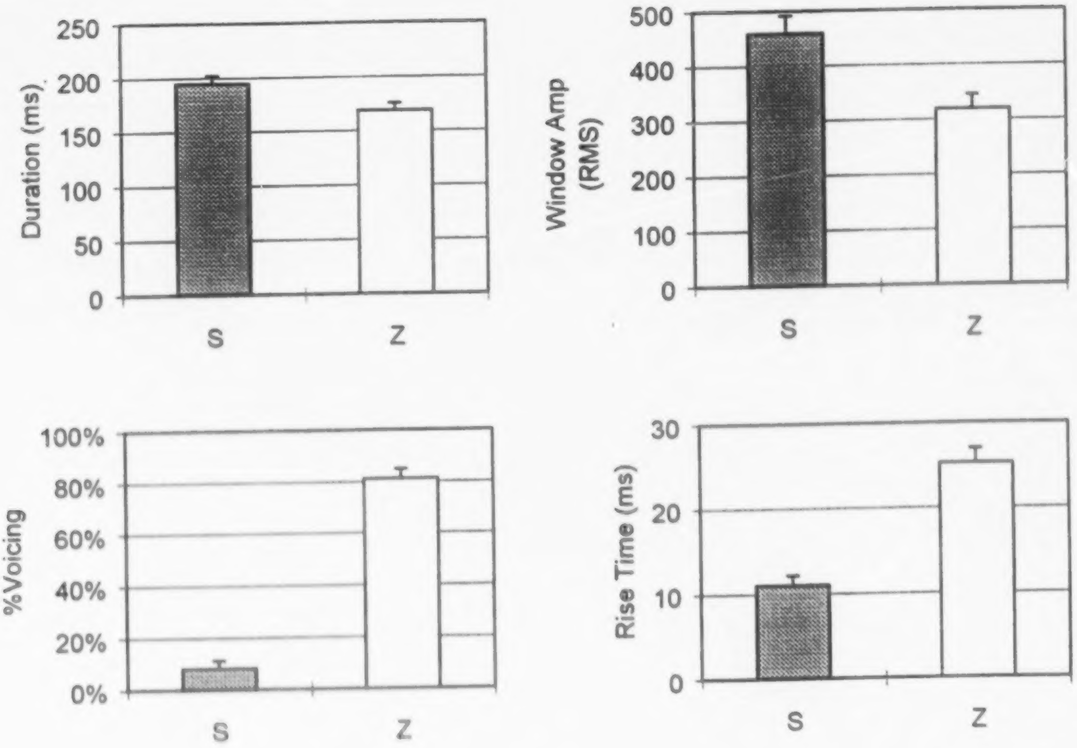


Figure 2. Mean and standard error for DURATION, WINDOW AMPLITUDE, %VOICING, and VOWEL RISE TIME for all productions.

Error Data

Categorical Errors

Two intended productions of [s] were [z]-like on every measure, and therefore appear to be categorical switches from [s] to [z]. Both of these productions were scored as errors in Frisch (1996). Acoustic characteristics for the categorical errors and the productions deemed to be completely normal are shown in Figure 3. The symbol indicates the intended production, [s] or [z]. The errors are marked by boxes around their symbols.

Insert Figure 3 about here

Voicing Errors

Several productions of [s] and [z] had abnormal %VOICING and/or VOWEL RISE TIME. However, they have normal WINDOW AMPLITUDE and DURATION. These errors are shown in Figure 4. Intended [s] which were partially voiced, but transcribed as [s], are marked by boxes around their symbols. These productions contained voicing at the onset of the fricative, but did not overlap with the following vowel. The intended [s] which were mostly voiced and were transcribed as [z] in the coding of the speech error experiment are marked by diamonds around their symbols. These productions contained fricative noise overlapping with the following vowel.

Insert Figure 4 about here

There was one intended [z] which was almost completely devoiced, indicated by the arrow. This production was, however, transcribed as [z]. It is entirely voiceless except at its very end where there is fricative noise overlapping with the following vowel onset. Intended [z] which were mostly devoiced are marked by circles around their symbols. These productions were transcribed as [z] with some difficulty in the original error experiment, and judged by the authors to be the most ambiguous tokens produced by this participant. They were voiced primarily in their beginning portions. One of these productions was corrected by the speaker, suggesting that the speaker thought the production was anomalous. This token was transcribed as [z] in the experimental coding, however, and thus not scored as an error.

Amplitude and Duration Errors

Several productions of [s] and [z] had abnormal WINDOW AMPLITUDE and/or DURATION, but normal %VOICING and VOWEL RISE TIME. Figure 5 shows amplitude and duration errors. Intended [s] with [z]-like duration and amplitude are marked by boxes around their symbols. One of these productions was transcribed as [z], it had higher vowel rise time than most other intended [s] productions (19.9ms).

Insert Figure 5 about here

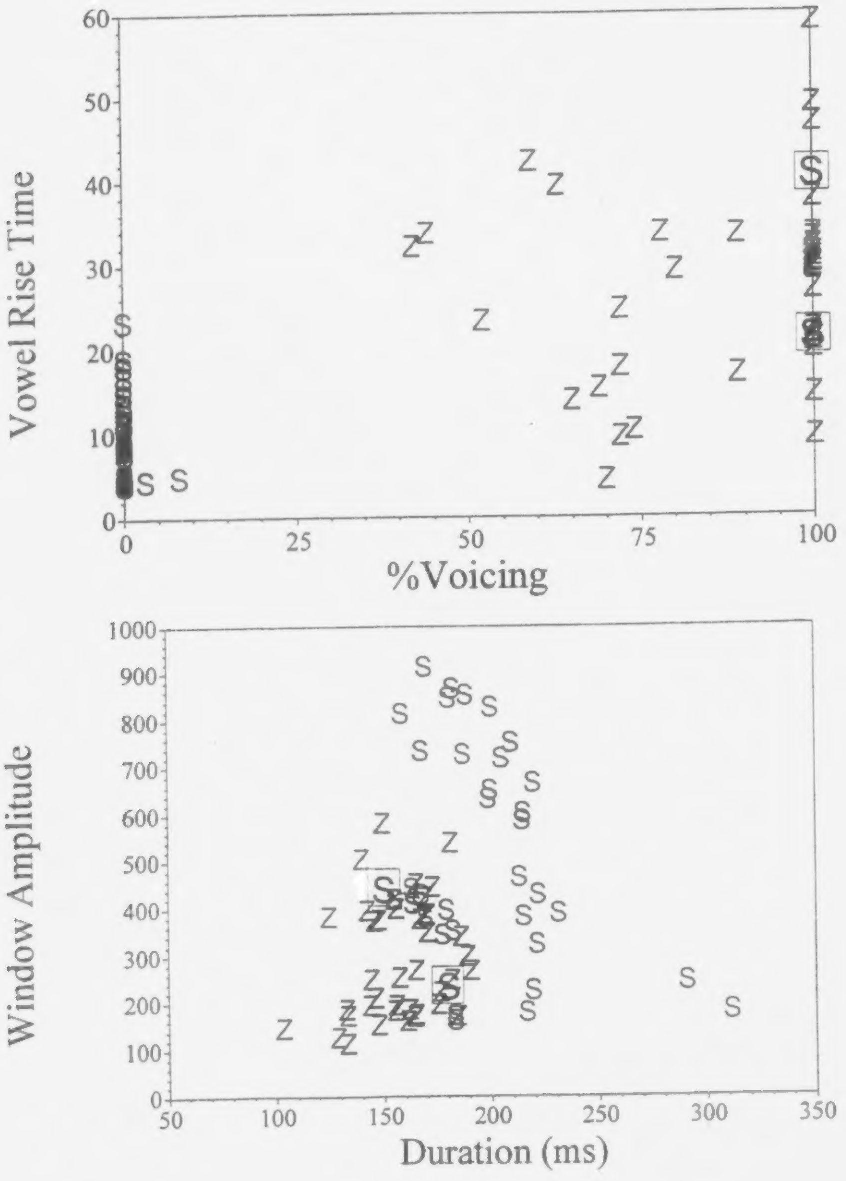


Figure 3. Categorical Errors.

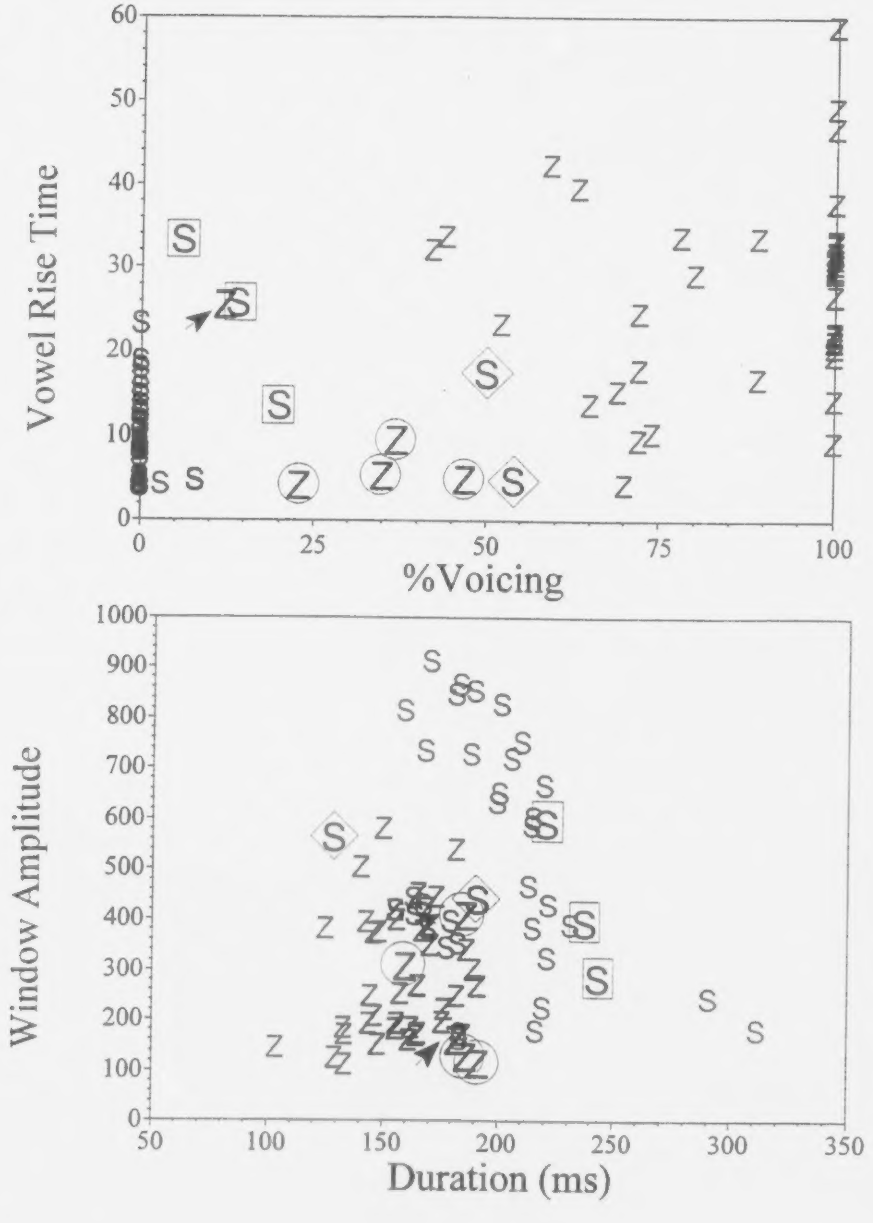


Figure 4. Voicing and Rise Time Errors.

BEST COPY AVAILABLE

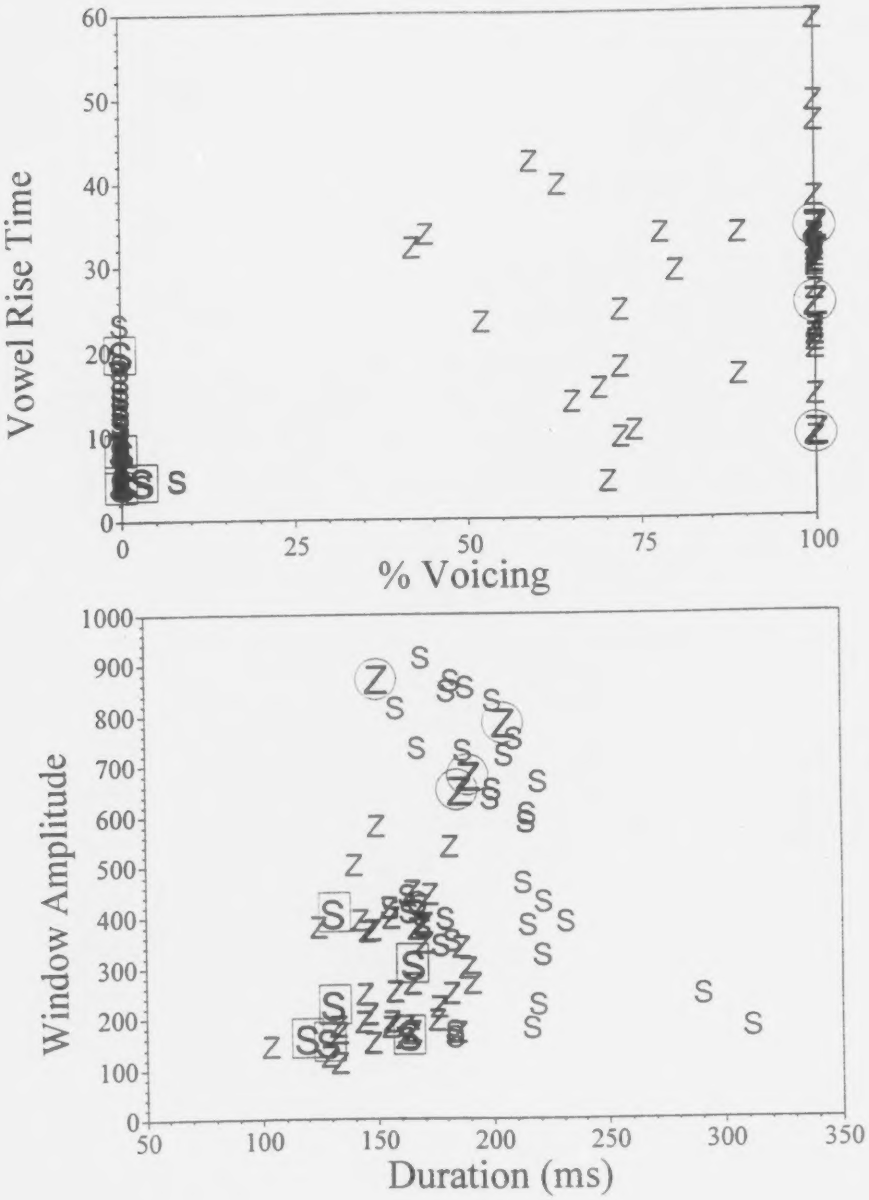


Figure 5. Frication Amplitude and Duration Errors.

The intended [z] with [s]-like amplitude are marked by circles around their symbols. These productions were all fully voiced, and were transcribed as [z] in the original experiment. Qualitatively, these sound to the authors like excellent examples of [z], regardless of their [s]-like duration. Examples such as these make it clear that some of the different acoustic cues for [s] and [z] are not perceptually equivalent.

Concatenated Errors

Two productions are extreme cases of errors which were corrected without hesitation, repetition, or cessation of the fricative noise. Waveforms for these productions are shown in Figure 6. As can be seen in the Figure, they have extremely long duration, appropriate for two independent productions. In both cases, the participant began with the incorrect production ([s] in the first case and [z] in the second) and ended with the correct production. In both cases, during the transition from the beginning to the end there was a period of reduced frication noise appropriate for [z], but with an absence of voicing. This is indicated in the figure as [z8].

Insert Figure 6 about here

Conclusion

Acoustic analysis of the speech errors in this study reveals acoustically categorical errors, presumably consistent with an error in a representational unit at the level of the segment or gestural constellation (Browman & Goldstein, 1986). However, a number of sub-featural errors were found. These errors may be the result of a single gestural error in line with the findings of Mowrey and MacKay (1990). Additionally, we have shown that sub-featural errors may or may not be auditorily contrastive, even if they are acoustically erroneous.

Instrumental investigation of speech errors shows that sub-lexical errors lie on a continuum between categorical and sub-featural errors. This pattern of errors can be revealed only by instrumental methods, as the errors are often not phonologically contrastive, and may not even be auditorily detectable. Individual gestures (as reflected in their acoustic consequences) do not obey constraints on phonotactic grammaticality (Mowrey & MacKay, 1990). This fact is most clearly apparent in the concatenation errors, which effectively produced [sz] and [zs] clusters word initially, a pattern not found in English. These violations were potentially sub-featural errors, however, so it may be the case that phonotactic regularity is upheld at the level of the segment. In other words, it may be the case that categorical segment errors do result in phonotactically regular words.

An analogous situation has been observed at other levels of language processing. Garrett (1975) found increased violations of the syntactic class constraint for units smaller than the word. Thus, the syntactic class constraint (e.g., nouns replace nouns) is upheld for words, but sometimes violated for morphemes, and frequently violated for segments. Dell and Gupta (1997) found the syllable position constraint (Boomer & Laver, 1968) is gradiently violated depending on the domain of the syllable position generalization. In general, for example, syllable onsets interact with onsets, obeying their syllabic affiliation. Dell & Gupta found that this law is sometimes violated within the scope of an utterance, but if there is a distributional constraint within the language (e.g., /h/ only appears as an onset, /N/ only appears

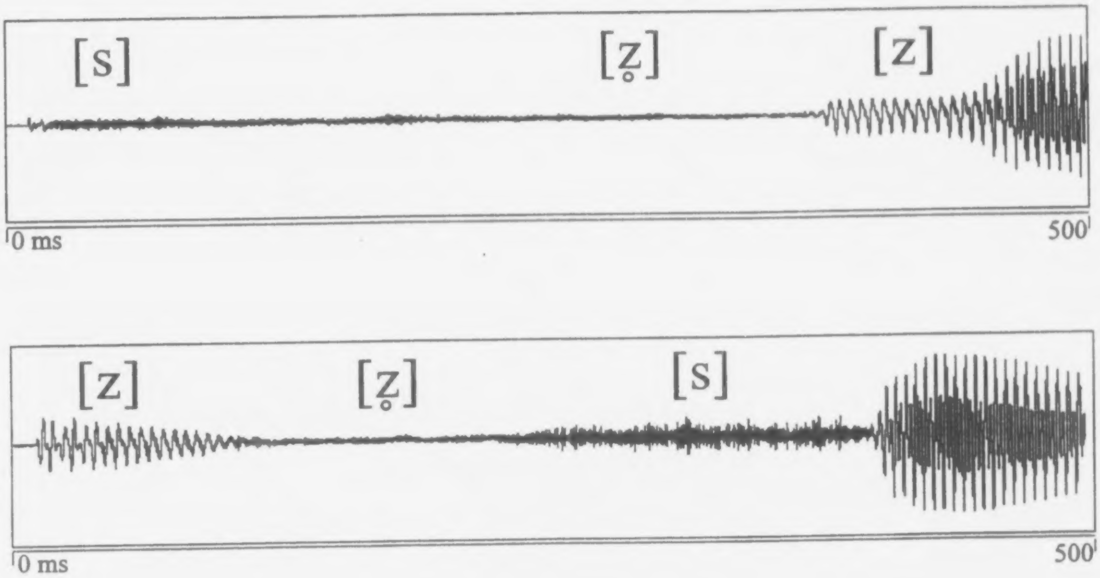


Figure 6. Concatenation Errors: [sz] with [z] intended (top) and [zs] with [s] intended (bottom).

as a coda in English) the syllable position constraint is never violated by those segments. Interestingly, they found that if the experiment is designed so that there is a distributional constraint within the experimental stimuli which is not found in the ambient language, that constraint is obeyed, and fewer violations of the syllable position constraint are found which also violate the distributional constraint within the experiment. Thus, they demonstrated gradient degrees of violation of the syllable position constraint, depending on the generality of the constraint in the participant's linguistic experience.

Future Work

Our current research plan is to combine the acoustic measurements presented in this study with analogous measurements for five other participants. Our long term goal is to find quantitative, rather than qualitative, evidence for or against a segmental level of organization in speech production. With the combined data, we plan to examine the statistical distribution of productions on each of the four dimensions presented here for [s] and [z]. If speech errors are solely the result of individual muscle mis-articulations (as proposed by Mowrey & MacKay) then we would expect to find distributions with a single mode for each dimension for each speaker for each consonant. Categorical errors would be instances where, by random variation, values of all dimensions co-occur in a single production which are appropriate for the other consonant. If, on the other hand, there is a segmental unit of organization which coordinates several gestures, then we expect to find a disproportionate number of categorical errors where all dimensions take extreme values simultaneously. So extreme values across dimensions will be correlated, and the distributions on each dimension for each speaker for each consonant will be bimodal. Since there is a great deal of variation over a relatively small sample of tokens for each speaker, we are also investigating statistical methods for combining data across speakers. Some patterns may not be reliably identified in every speaker, in which case group data may provide insight that an individual speaker analysis, such as the one presented here, might miss.

We also plan to investigate the perceptual side of speech errors, using the corpus of errors studied here in a series of playback experiments. Our measurements provide a quantitative scale on which errors can occur to different degrees. The degree to which errors of different degrees along different dimensions can be detected by naive listeners will be tested using tokens from this corpus. We are interested both in the reliability of error detection across listeners as well as individual differences in perceptual boundaries between listeners. These results bear directly on the reliability of speech error data collected in the laboratory or opportunistically.

References

Boomer, D. & Laver, J. (1968). Slips of the tongue. *British Journal of Disorders of Communications*, 3, 1-12.

Boucher, V. (1994). Alphabet-related biases in psycholinguistic enquiries: considerations for direct theories of speech production and perception. *Journal of Phonetics*, 22, 1-18.

Browman, C. & Goldstein, L. (1986). Towards an articulatory phonology. *Phonology Yearbook*, 3, 219-252.

Dell, G. & Gupta, P. (1997). Producing and representing serial order. Paper presented at the Carnegie Symposium on Emergentist Approaches to Language, May 1997, Pittsburgh, PA.

Frisch, S. (1996). *Similarity and frequency in phonology*. Unpublished Ph.D. Dissertation, Northwestern University, Evanston, IL.

Fromkin, V. (1971). The non-anomalous nature of anomalous utterances. *Language*, 47(1), 27- 52.

Garrett, M. (1975). The analysis of sentence production. In G. Bower (ed.), *The Psychology of Learning and Motivation* (pp. 133-177). New York: Academic Press.

Klatt, D. (1976). Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *Journal of the Acoustical Society of America*, 59, 1208-1221.

Laver, J. (1979). Slips of the tongue as neuromuscular evidence for a model of speech production. In H. Dechert & M. Raupach (eds.), *Temporal variables in speech* (pp. 21-26). The Hague: Mouton.

Mowrey, R. & MacKay, I. (1990). Phonological primitives: Electromyographic speech error evidence. *Journal of the Acoustical Society of America*, 88(3), 1299-1312.

Pickett, J. (1980). *The sounds of speech communication*. Baltimore, MD: University Park Press.

Shattuck-Hufnagel, S. (1992). The role of word structure in segmental serial ordering. *Cognition*, 42: 213-259.

Strevens, P. (1960). Spectra of fricative noise in human speech. *Language and Speech*, 3, 32-49.

Wells, R. (1951). Predicting slips of the tongue. *Yale Scientific Magazine*, December, 9-12.

Wickelgren, W. (1965). Distinctive features and errors in short-term memory for English vowels. *Journal of the Acoustical Society of America*, 38, 583-588.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 21 (1996-1997)
Indiana University

Experimental Evidence for Abstract Phonotactic Constraints[1]

Stefan Frisch and Bushra Zawaydeh[2]

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405

---

[1] This work supported by NIH-NIDCD Training Grant DC00012 to Indiana University.

[2] Department of Linguistics at Indiana University, Bloomington, IN.

Experimental Evidence for Abstract Phonotactic Constraints

Abstract. This paper provides evidence for the psychological reality of a highly abstract phonotactic constraint within the verbal roots of Arabic, known as OCP-Place. In a novel root rating task, non-roots containing constraint violations were rated less acceptable than control non-violations. Ratings were also influenced by the lexical neighborhood density of the non-roots within the lexicon of occurring roots. Lexical characteristics of the non-root stimuli were controlled so that the difference between constraint violations and controls was not in the type frequency of consonants and consonant pairs involved, but instead a difference between a linguistically systematic and an accidental gap in the lexicon. In other words, the abstract constraint is a psychologically real factor in judging non-word acceptability.

Introduction

Generative and connectionist approaches to linguistic knowledge make different claims about the psychological reality of linguistic constraints. In distributed connectionist models, constraints are emergent properties of the set of lexical items and do not exist as independent mental entities. In formal generative grammar, constraints (or rules) are the core of linguistic knowledge and knowledge about individual lexical items is restricted to idiosyncratic information. This paper provides evidence based on a pilot experiment for the psychological reality of a highly abstract phonotactic constraint within the Arabic verbal roots. Our results are consistent with the generative approach, since we find evidence for an abstract phonotactic constraint which is independent of specific patterns in the lexicon. However, we find that lexical patterns influence acceptability judgments as well, as is predicted by a connectionist account.

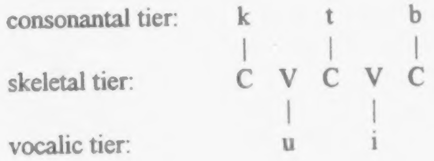
In addition, our results are theoretically important because that we find lexical factors which have previously been examined only in English to have an influence in Arabic as well. In particular, we found that lexical neighborhood density (Goldinger, Luce, & Pisoni, 1989; Luce, 1986) influences acceptability ratings of non-words in Arabic.

Introduction to OCP-Place in Arabic

Morphology

The verbal roots of Arabic have a root-and-pattern morphology, and the underlying form of Arabic verbal roots is assumed to be a sequence of consonants, e.g., /k t b/. The canonical root contains three consonants. Vowels and a syllabic structure are provided by separate morphemes. This situation provides a rationale for segregating the consonants and vowels into separate autosegmental tiers (McCarthy, 1979), as in now famous examples such as kutib 'to be written', shown in (1).

(1)



Consonant Inventory

Arabic has a large consonant inventory, which includes nearly all of the consonants of English, as well as two series of consonants not found in English. The complete inventory is given in (2). The EMPHATIC consonants /T, D, S, Z/ are similar to the familiar English consonants /t, d, s, z/ but they are articulated with an additional constriction in the pharynx. The GUTTURAL consonants /χ, ʁ, ħ, ʕ, h, ʔ/ have place of articulation in the back of the throat, ranging from the uvula to the larynx. Their manner of articulation is considered to be APPROXIMANT by Catford's (1977) definition. The voiceless gutterals have a turbulent noise source which is absent in the voiced gutturals (see McCarthy, 1994, for an excellent summary of the phonetic character of the Arabic gutturals).

(2)	Labial	Coronal	Emphatic	Velar	Uvular	Pharyngeal	Laryngeal
		t	θ	k	q		ʔ
	b	d	ð	g			
	f	θ, s	ʃ		χ	ħ	h
		ð, z	ʒ		ʁ	ʕ	
		ʃ					
		l, r					
	m	n					
		w, y					

For most Arabic speakers the underlying velar /g/ is realized as [g], [dZ], or [Z]. However, for the purposes of the phonotactic constraint as it is reflected in the lexicon, this consonant patterns as a velar (McCarthy 1994). In the Levantine dialect which our participants speak, it is realized as [Z]. The possibility that this consonant would pattern differently in a behavioral experiment is not addressed here, but is an interesting topic for future research.

Phonotactics

Arabic has a well-known phonotactic constraint that restricts the set of allowable consonant sequences in roots based on place of articulation: Arabic roots rarely contain homorganic consonant pairs (Greenberg, 1950). This has traditionally been analyzed as a co-occurrence constraint, known as OCP-PLACE (McCarthy, 1986, 1988, 1994), which prohibits repeated place features within a verbal root.

The traditional approach to these co-occurrence restrictions is to divide the Arabic consonants into natural classes, with co-occurrence constraints applying within these classes. The major co-occurrence classes discussed by Greenberg and McCarthy are presented in (3). In their analyses, consonants in any one of these classes are claimed to co-occur freely with consonants from any other class, and within any class consonants tend not to co-occur, with two exceptions. First, the velars (3c) cannot co-occur with the uvular approximants {χ, ʁ}, though they can co-occur with the other gutturals. Second, among the coronal obstruents, there are far more roots containing one fricative and one stop than roots containing two fricatives or two stops. The phonological status of the glides {w, y} is unclear and they are typically excluded from analyses of OCP-Place. There may be a co-occurrence restriction between the glides and there is no evidence for a co-occurrence restriction between the glides and any other consonants (McCarthy 1994).

Major co-occurrence classes:

(3) a. Labials = {b, f, m}
b. Coronal Obstruents = {t, d, T, D, θ, ð, s, z, S, Z, ʃ}
c. Velars = {k, g, q}
d. Gutturals = {χ, ʁ, ħ, ʕ, h, ʔ}
e. Coronal Sonorants = {l, r, n}

Pierrehumbert (1993) demonstrated that the categorical statement of OCP-Place, based on the major co-occurrence classes, is incorrect. She showed that OCP-Place is gradient, and sensitive to the similarity of the homorganic consonants that are involved. For example, roots containing repeated identical consonants are never found (McCarthy, 1986), but roots containing homorganic stops and fricatives, e.g., /d s w/, are well attested.

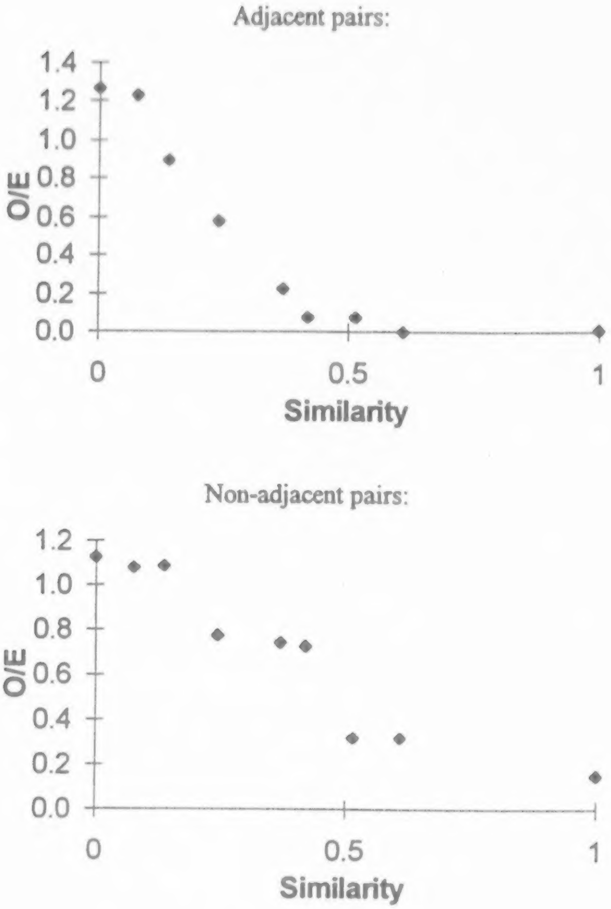
More recently, Frisch, Broe, and Pierrehumbert (1997) developed a quantitative model of co-occurrence in Arabic based on a novel similarity metric for consonants. They analyzed the Arabic verb lexicon as approximated by a dictionary (Cowan, 1979; see Oldfield, 1966), categorizing consonants by similarity rather than major class. They expressed degrees of co-occurrence of consonants using the ratio of the observed number of consonant pairs in the dictionary (O) compared to the number which would be expected if consonants combined to form roots at random (E). Random combination was computed by multiplying the probabilities of consonant occurrence in each position in the root. For example, the expected frequency of the root /d s w/ is the product of the probability of /d/ in the first position times the probability of /s/ in second position times the probability of /w/ in third position times the total number of roots (2674 in Cowan, 1979).

Figure 1 shows the co-occurrence pattern (O/E) as a function of similarity for individual consonant pairs. The top figure is aggregate O/E for 'adjacent' consonant pairs (combinations of C1C2 or C2C3) and the bottom figure is aggregate O/E for 'non-adjacent' consonant pairs (combinations of C1C3). Both figures reflect the dependence of the co-occurrence constraint on similarity. As similarity increases, relatively fewer roots are found which contain consonant pairs of that similarity.

Insert Figure 1 about here

Note also that the constraint is weaker for non-adjacent consonants. Pierrehumbert (1993) explains the weakening of the constraint on interference in the perception of similarity for the more distant non-adjacent consonants (cf. Ericksen & Shultz, 1979; Massaro, 1970; Pisoni, 1973). For example, in the hypothetical root */d m t/, the medial /m/ provides interference in determining the similarity of /d/ to /t/. For more discussion of the influence of cognitive factors on phonology, see Frisch (1996). The non-categorical nature of the pattern in the lexicon is clearly evident in the gradient and cumulative effects of similarity and distance.

The studies of Greenberg (1950), McCarthy (1986, 1988, 1994), Pierrehumbert (1993), and Frisch et al (1997) are all based on the statistical analysis of a dictionary as an approximation of a native speaker's lexical knowledge. They are therefore suspect on the grounds that the distribution of words in the dictionary is influenced by historical ancestry, and may not be a productive part of the grammar. Thus, the



Adjacent pairs:

Non-adjacent pairs:

Figure 1. Aggregate O/E for adjacent consonant pairs (top panel) and non-adjacent consonant pairs (bottom panel) in the lexicon of Arabic verbal roots as a function of consonant pair similarity.

OCP-Place constraint may not be part of the synchronic grammar of native speakers. We conducted an experiment to test whether the characterization of the OCP-Place constraint from the dictionary studies accurately reflects the native speaker's implicit knowledge. This experiment also examined whether knowledge of the abstract constraint exists independently of the pattern of lexical items in the dictionary. In other words, are violations of the constraint less word-like than equally rare, but not phonologically regular, gaps in the lexicon.

Experiment

Previous research on acceptability judgments for non-words has found that acceptability ratings are influenced by the how much the non-word is like existing forms in the language (Greenberg & Jenkins, 1962; Ohala & Ohala, 1986). Recent work has quantified word-likeness using the expected frequency of the phonemes in the word. Acceptability ratings for CVC non-words in English are affected by the frequency of the VC combination in the lexicon (Treiman, Kessler, Knewasser, Tincoff, & Bowman, 1996). Vitevitch, Luce, Charles-Luce, & Kemmerer (1997) found analogous effects for consonant frequency in bisyllabic CVC.CVC sequences. Coleman & Pierrehumbert (1997) found the logarithm of the expected frequency of the entire word (where expected frequency was determined by independent combination of the positionally correct probabilities of the onsets and rimes) to be a good predictor of acceptability for a variety of English non-words. These studies all used the type or token frequency of the units which comprise the non-word as a measure of the expected probability of the non-word.

There is a second approach to word-likeness in the spoken word recognition literature which is based more directly on the similarity of a non-word to the set of existing words. The SIMILARITY NEIGHBORHOOD of a target word is the set of words which differ from the target word by at most one phoneme (either substituted, added, or deleted). The NEIGHBORHOOD DENSITY is the size of that set (Goldinger, Luce, & Pisoni, 1989). Luce (1986) and Luce & Pisoni (1998) found that the neighborhood density of a word/non-word influences accuracy and reaction time for a variety of perceptual tasks in English, including lexical decision. In an auditory lexical decision task, a participant must respond as quickly as possible to whether the stimulus is a word or not. Luce (1986) found that non-words in high density neighborhoods were less accurately identified as non-words than non-words in low density neighborhoods. He also found that reaction times for responses were longer for non-words in high density neighborhoods than for non-words in low density neighborhoods.

In order to test the influence of the OCP-Place constraint on acceptability, factors were controlled which influence acceptability and which are unrelated to the OCP constraint. The experimental stimuli were designed to match OCP violations with non-violations that were phonotactically equally probable and had the same number of lexical neighbors.

Stimulus Materials

All non-words in the experiment were non-existing three consonant verb roots. The set of 2764 existing three consonant roots from Wehr's dictionary of Arabic (Cowan, 1979) were used to compute lexical statistics for the non-word stimuli. This is the same dictionary which was used in the studies of Pierrehumbert (1993) and Frisch et al (1997). The stimuli were divided into two sets.

Stimulus Set 1 was designed for a three-way factorial ANOVA with expected frequency, neighborhood density, and OCP-Place violation as factors. The expected (type) frequency of a non-occurring three consonant sequence was computed by independent combination of the three consonants, as in Pierrehumbert (1993). For example, the expected frequency of */ʔ f z/ is

Expected(*/ʔ f z/) = P(/ʔ C C/) × P(/C f C/) × P(/C C z/) × 2674

where P(/x y z/) is the probability of the sequence in the lexicon, and C is any consonant. In other words, the expected frequency is based on the product of the positionally correct phoneme probabilities.

Neighborhood density was computed by single phoneme substitution. In other words, the set of neighbors for a verbal root was taken to be the set of roots which differ by only one consonant. For example, the neighborhood density of */ʔ f z/ is

Density(*/ʔ f z/) = N(/ʔ f C/) + N(/C f z/) + N(/ʔ C z/)

where N(/x y z/) is the number of roots with the sequence in the lexicon, and C is any consonant. The density is equal to the number of roots which share two of the three consonants in the target root, and thus approximates the phonemic similarity of the target root to other roots.

It should not be surprising that expected frequency and neighborhood density are correlated over the entire lexicon. If a target root has many neighbors, then there are many other roots which contain the consonants in the target root and the expected frequency of the target root is relatively high. It is only possible to independently vary expected frequency and neighborhood density if the expected frequency and density is relatively low. Since OCP-Place violations tend not to occur, non-words containing OCP-Place violations are also of relatively low neighborhood density. In addition, expected frequency is correlated with word-hood (Pierrehumbert 1994, Frisch 1996), so non-words tend to have lower expected frequency than words. The stimuli were, on the whole, improbable sequences with few neighbors. The stimuli were grouped into two categories for each factor, as shown in Table 1 (mean values are shown in parenthesis). There were 160 total stimuli, 20 per three-way combination of categories.

Table 1

Three way categorization of stimuli in Stimulus Set 1.

Frequency Category	Exp. Frequency	Density Category	Density	OCP Category
High Freq	0.12-0.25 (0.171)	Dense	11-20 (13.2)	OCP Violation
Low Freq	0.06-0.12 (0.068)	Sparse	1-10 (8.2)	No OCP Violation

Stimulus Set 2 was designed to test whether Arabic speakers' acceptability ratings differentiate SYSTEMATIC GAPS (due to the OCP-Place constraint) and ACCIDENTAL GAPS (chance gaps in consonant sequences with no systematic description). In this stimulus set, OCP-Place violations and control stimuli were balanced for expected frequency and neighborhood density as for Stimulus Set 1. In addition, the stimuli were balanced for the transitional frequency of their consonant pairs. Since OCP-Place violations tend not to occur, violating pairs are underrepresented in the lexicon. In each control, one consonant pair was highly underrepresented in the lexicon of Arabic, but did not form a natural class of examples with other underrepresented pairs. In other words, the controls contained accidental gaps while the OCP-Place violations were systematic.

OCP-Place violations were matched with non-violations that had the same transitional frequency of the C1C2, C2C3, and C1C3 consonant pairs. For example, for the non-existing root *⁄m g t/ there are 3 roots of the form /m g C/, no roots of the form /C g t/ and two roots of the form /m C t/, where C is any consonant. The corresponding stimulus with an OCP-Place violation is /b S T/, which contains the emphatic coronal obstruent pair /S, T/. There are 3 roots of the form /b S C/, no roots of the form /C S T/ and two roots of the form /b C T/, where C is any consonant. There were 40 total stimuli, 20 violations and 20 non-violations.

Participants

Five native speakers of Levantine Arabic (including the second author) who are students at Indiana University, participated in the experiment. They were paid for their participation.

Methods

Stimuli were presented orthographically in Arabic. Stimuli were presented in infinitival form. The participants rated the non-words on a 1-7 scale (1 = impossible, 7 = sounds just like a verb of Arabic). All 200 non-words were rated in a single session, which took approximately 30 minutes. The verbs were presented in pseudo-random order, so that no two stimuli from the same category in either stimulus set appeared in sequence. Items from Stimulus Set 1 were randomly mixed with items from Stimulus Set 2.

Results: Stimulus Set 1

Analysis of variance of participants' ratings of the items in Stimulus Set 1 showed an extremely strong main effect for OCP-Place violation ($F = 79.1$, $p < 0.0001$). OCP-Place violations were rated much less acceptable than non-violations by all participants. There was also a main effect of density class ($F = 7.3$, $p = 0.007$). Non-words with higher neighborhood density were rated as more acceptable than non-words with lower neighborhood density by all participants. Frequency class was not a significant factor ($F < 1$), although the overall trend was in the expected direction. Non-words with higher expected frequency were rated slightly more acceptable than non-words with lower expected frequency.

There was also a significant interaction between density class and OCP-Place violation ($F = 4.6$, $p = 0.035$). For OCP-Place violations, the effect of density was reduced. This may have been a floor effect, because many of the OCP-Place violations were given the minimum rating of 1. The main effects of density and OCP-Place violation, as well as the interaction between the two, is shown in Figure 2. Expected frequency categories were collapsed in this figure since no effect of expected frequency was found.

Insert Figure 2 about here

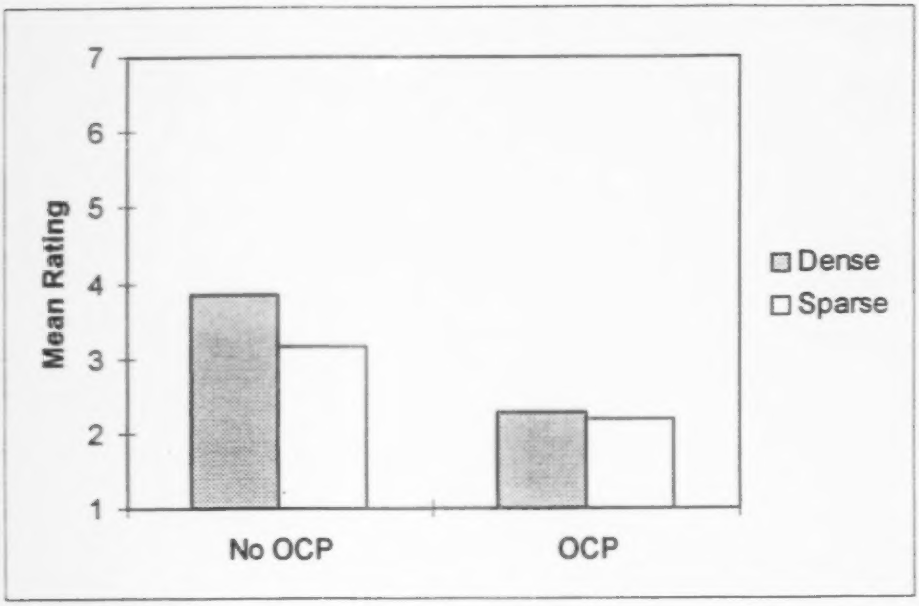


Figure 2. Mean ratings of non-violations and OCP violations by neighborhood density group. Violations were rated much less acceptable than non-violations. Lexical neighborhood density had a greater effect for non-violations. Non-violations from dense neighborhoods were rated much more acceptable than non-violations from sparse neighborhoods.

Results: Stimulus Set 2

Analysis of variance of participants' ratings with OCP-Place violation and participant as factors showed a main effect of OCP-Place violation (F = 14.2, p = 0.0002). OCP-Place violations were rated as much less acceptable than non-violations even though the non-violations were gaps in the Arabic lexicon. Mean ratings for each participant are shown in Figure 3. All five participants judged OCP-Place violations to be less acceptable than the control non-violations which contained accidental gaps.

Insert Figure 3 about here

Discussion

This experiment found lexical neighborhood density to have an effect on acceptability ratings. Previously, lexical neighborhood effects have only been observed in perceptual tasks in English. We have shown that neighborhood density has an effect on acceptability judgments, which is a linguistic task. Further, neighborhood density is a relevant characteristic of the lexicon of consonantal roots of Arabic. The Arabic lexicon is conceptually more abstract than the English lexicon, as the consonantal sequences are extracted from word forms containing both consonants and vowels, and some of the consonants in the word forms are not part of the root.

The much lower ratings for OCP violations versus non violations in Stimulus Set 1 provide some evidence for a psychologically real OCP-Place constraint in the grammar of Arabic speakers. When lexical factors like expected frequency of the consonant sequence in the verb and the number of phonological neighbors of the verb are taken into account, there is still a strong effect of violating the OCP-Place constraint on acceptability. However, while the stimuli were balanced for the overall neighborhood density of the words and the expected frequency of the phoneme combinations, there were still some differences in lexical characteristics between violations and non-violations in Stimulus Set 1. In particular, the violations tended to have one consonant pair, the pair which violated the OCP, which did not occur in other words in the lexicon. The neighborhood density was still balanced as the other consonant pairs in the word occurred more frequently. For the non-violations, this was generally not the case. Thus, based only on the results for Stimulus Set 1, it is possible that acceptability is not based on an abstract constraint, but rather on the transitional probability of the consonant sequences in the root.

Stimulus set 2 compared consonant pairs with equal transitional probability, half of which were accidental gaps, and half of which were systematic gaps (i.e. OCP-Place violations). The effects of the OCP-Place constraint were found even when lexical gaps were compared. Participants rated systematic gaps much worse than accidental gaps. Systematic gaps, which are the basis for linguistic constraints, have independent influence on acceptability judgments from lexical characteristics.

Conclusions and Future Directions

The present findings demonstrate that non-words which contain constraint violations are judged less acceptable than control non-violations. In addition, the difference between constraint violations and controls was not in the type frequency of consonants and consonant pairs involved, but instead was due to a difference between a linguistically systematic and an accidental gap in the lexicon. In other words, the constraint is a psychologically real abstraction and not only a reflection of the similarity space of lexical

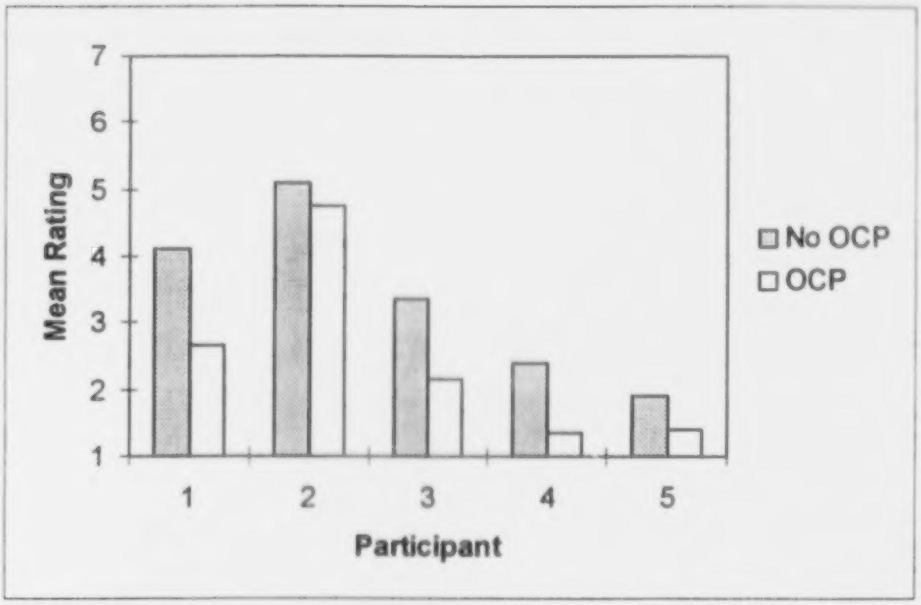


Figure 3. Mean ratings for each participant of stimuli with accidental (No OCP) or systematic (OCP) gaps. Systematic gaps were rated much less acceptable by all participants.

types. The existence of an abstract constraint supports the generative notion of linguistic knowledge as containing constraints abstracted from their lexical contexts.

Note, however, that these results are not inconsistent with connectionist or emergent approaches to language. What has been shown here is that Arabic speakers do know an abstract generalization about their language. This generalization is emergent from the pattern of lexical items (the generalization was observed by linguists examining the patterns over a dictionary, after all), and thus could potentially be learned by a connectionist system and need not be assumed to be innate. The point to be understood is that the system would have to be able to learn generalizations to the degree of abstraction generally assumed in linguistic theory.

We are currently analyzing data from the full version of this experiment, which used 30 participants living in Amman, Jordan. Preliminary results indicate that the OCP-Place effects are just as strong for these participants as they are for those who participated in the pilot experiment described earlier, and that these participants also differentiated accidental and systematic gaps just as reliably.

References

Catford, J. C. (1977). *Experimental problems in phonetics*. Edinburgh: Edinburgh University Press.

Coleman, J. and Pierrehumbert, J. (1997). Stochastic phonological grammars and acceptability. Paper presented the Third Meeting of the ACL Special Interest Group in Computational Phonology, Madrid.

Cowan, J. (1979). *Hans Wehr: a dictionary of modern written Arabic*. Wiesbaden, Germany: Otto Harrasowitz.

Eriksen, C. and Schultz, D. (1978). Temporal factors in visual information processing: a tutorial review. In J. Requin (ed.), *Attention and Performance VII*, Earlbaum, Hillsdale, NJ.

Frisch, S. (1996). *Similarity and frequency in phonology*. Unpublished Ph.D. dissertation, Northwestern University.

Frisch, S., Broe, M., and Pierrehumbert, J. (1997). Similarity and phonotactics in Arabic. Manuscript, Indiana University and Northwestern University. Submitted for publication.

Greenberg, J. (1950). The patterning of root morphemes in Semitic. *Word*, 5, 162-181.

Greenberg, J. and Jenkins, C. (1964). Studies in the psychological correlates of the sound system of American English. *Word*, 20, 157-177.

Goldinger, S., Luce, P., and Pisoni, D. B. (1989). Priming lexical neighbors of spoken words: effects of competition and inhibition. *Journal of Memory and Language*, 28, 501-518.

Luce, P. (1986). *Neighborhoods of words in the mental lexicon*. (Research on Speech Perception Technical Report No. 6). Bloomington, IN: Speech Research Laboratory, Indiana University.

Luce, P. & Pisoni, D. B. (1998). Recognizing spoken words: The Neighborhood Activation Model. *Ear and Hearing*, 19, 1-36.

Massaro, Dominic. (1970). Retroactive interference in short-term recognition memory for pitch. *Journal of Experimental Psychology*, 83, 32-39.

McCarthy, J. (1979). *Formal problems in Semitic phonology and morphology*. New York: Garland.

McCarthy, J. (1986). OCP effects: gemination and antigemination. *Linguistic Inquiry*, 172, 207-263.

McCarthy, J. (1988). Feature geometry and dependency: a review. *Phonetica*, 43, 84-108.

McCarthy, J. (1994). The phonetics and phonology of Semitic pharyngeals. In P. Keating (ed.), *Papers in laboratory phonology III* (pp.191-283). Cambridge: Cambridge University Press

Ohala, J. and Ohala, M. (1986). Testing hypotheses regarding the psychological manifestation of morpheme structure constraints. In J. Ohala and J. Jaeger (eds.), *Experimental phonology* (pp. 239-252). New York: Academic Press.

Oldfield, R. C. (1966). Things, words, and the brain. *Quarterly Journal of Experimental Psychology*, 18, 340-353.

Pierrehumbert, J. (1993). Dissimilarity in the Arabic verbal roots. *Proceedings of the North East Linguistics Society*, 23, 367-381.

Pierrehumbert, J. (1994). Syllable structure and word structure. In P. Keating (ed.) *Papers in laboratory phonology III* (pp. 168-188). Cambridge: Cambridge University Press.

Pisoni, D. B. (1973). Auditory and phonetic memory codes in the discrimination of consonants and vowels. In *Perception and Psychophysics*, 13, 253-260.

Treiman, R., Kessler, B., Knewasser, S., and Tinkoff, R. (1996). Adults' sensitivity to phonotactic probabilities in English words. Manuscript submitted for publication, Wayne State University, Detroit, MI.

Vitevitch, M., Luce, P., Charles-Luce, J., and Kemmerer, D. (1997). Phonotactics and syllable stress: implications for the processing of nonsense words. *Language and Speech*, 40, 47-62.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 21 (1996-1997)
Indiana University

Experience with Sinewave Speech and the Recognition of Sinewave Voices[1]

Sonya M. Sheffert, David B. Pisoni, Nathan R. Large,
Jennifer M. Fellowes[2] and Robert E. Remez[2]

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405

---

[1] This research was supported by NIH-NIDCD Research Grant DC00111 and NIH-NIDCD Training Grant DC00012 to Indiana University, Bloomington and NIH-NIDCD Research Grant DC00308 to Barnard College, Columbia University, New York.

[2] Department of Psychology, Barnard College, New York, NY.

Experience with Sinewave Speech and the Recognition of Sinewave Voices

Abstract. This study explores the learning and generalization of voices using sinewave speech patterns. The present experiment was designed to address an asymmetry in generalization performance reported in a previous perceptual learning study (Sheffert, Pisoni, Fellowes & Remez, 1996). In that study, listeners who became familiar with a talker's voice through listening to sinewave replicas of naturally produced sentences showed excellent generalization to both novel sinewave and natural speech sentences, whereas speaker-specific knowledge acquired through natural speech only generalized to natural speech sentences. The primary difference between the two experiments was the amount of experience subjects had with sinewave materials before the generalization tests. The present experiment tested the hypothesis that the ability to recognize familiar voices from sinewaves depends on having prior phonetic experience with the unusual acoustic properties of sinewave signals. The experiment consisted of three phases: Training, transcription, and generalization. In the training phase, listeners learned to categorize ten talkers from naturally produced sentences. In the transcription phase, listeners were familiarized with sinewaves by transcribing sinewave sentences produced in an unfamiliar voice. In the generalization phase, listeners' ability to recognize the ten talkers from a novel naturally produced sentences and sinewave replicas was assessed. The results confirmed the earlier findings of Sheffert et al. (1996) by demonstrating that speaker-specific knowledge acquired during the perceptual training task generalized readily to novel natural utterance, but not to novel sinewave utterances. The data also show that prior exposure to the unusual nonspeech tonal patterns did not improve generalization performance. This pattern of results demonstrates that subjects' ability to exploit the talker-specific phonetic information present in the sinewave replicas does not depend on having phonetic experience with sinewave speech patterns.

Introduction

The present experiment is concerned with the perceptual learning and generalization of voice information. It is part of a series of experiments aimed at determining what properties of the acoustic signal support familiar voice recognition. This particular study is a control experiment designed to address an asymmetry in generalization performance found in two previous perceptual training experiments (Sheffert, Pisoni, Fellowes & Remez, 1996). In one study, sinewave replicas of naturally produced utterances were used to train listeners to identify 10 talkers. The sinewave signals were composed of three time-varying sinusoids that reproduced the center frequency and amplitude patterns of the oral, nasal and fricative resonances of the natural utterance. These nonspeech patterns preserved coarse-grained properties of the talker's vocal tract transfer function, including idiosyncratic phonetic variation, while eliminating cues to voice quality, such as fundamental frequency, harmonic structure and the fine-grained power spectra of nasals and vowels.

Surprisingly, subjects were able to learn to identify the ten talkers from these highly impoverished nonspeech sounds, although it took several training sessions. This result is important because it shows that perceptual learning of voices can be accomplished solely from the time-varying phonetic attributes encoded in the vocal tract transfer function. That is, the acoustic cues typically implicated in talker identification (Bricker, & Pruzansky, 1976; Laver, 1991) are not necessary for voice learning.

Our earlier experiment also showed that subjects who learned the voices from sinewave patterns were able to identify a talker from novel sinewaves and from novel natural speech samples with equal accuracy. This result indicates that the same acoustic correlates of voice identity were being utilized in both the sinewave and natural speech generalization tests, and supports the proposal that phonetic properties of an utterance jointly specify words and talkers (Remez, Fellowes & Rubin, 1997).

A very different pattern of generalization was found in a second experiment in which subjects learned to identify the talkers from natural utterances. The results showed that subjects quickly learned to identify a talker from naturally produced sentences, and that this knowledge generalized readily to novel natural speech utterances. However, speaker-specific knowledge did not generalize to novel sinewaves. In fact, the accuracy of identifying a speaker on the sinewave generalization test was only marginally above chance. Not only does this finding contrast with the generalization pattern in the first experiment, but it also contrasts with previous data from Remez et al. (1997).

A methodological difference that may account for the differences observed across the two training experiments is the amount of exposure subjects had to sinewave materials prior to the sinewave generalization test. In Experiment 1, subjects had several days of training with the sinewaves and were accustomed to the peculiar acoustic quality of the tonal analogs and were able to perceive the tonal analogs as speech-like. Similarly, the listeners used in Remez et al. (1997) were largely composed of speech scientists who had prior exposure to sinewave signals. In contrast, subjects in Experiment 2 had no prior experience with sinewaves. It is possible that the novelty of the sinewaves led subjects in Experiment 2 to focus on the unusual auditory impressions the tonal signals evoke, rather than on the talker-specific phonetic information present in the signals. Failing to attend to the transfer-relevant phonetic information in the sinewaves may have prevented the listeners from showing learning on the sinewave voice recognition test.

The present experiment tested the hypothesis that the ability to recognize familiar voices from sinewaves depends on having prior experience with sinewave signals. We attempted to resolve this issue by providing experience to our listeners using a sinewave transcription task. This task follows the natural speech perceptual training task but precedes the generalization tests. The purpose of the sinewave transcription task was to accustom listeners to the unusual timbre of the signals and to facilitate the perception of the phonetic information in the sinewaves. Subjects were provided with several repetitions of 29 sinewave replicas. None of these items duplicated any of the items used in the training or transcription tasks. In addition, the sentences were produced by a male voice who was not one of the talkers in the training or generalization phases of the experiment (R.E.R). Subjects were informed that the sinewaves were analogs of the natural speech tokens and were encouraged to listen to them as they would speech. The participants simply listened to each sentence and transcribed what they heard.

The general design of the experiment was identical to Experiment 2 in Sheffert, et al. (1996), with the exception of the transcription task. Specifically, the familiarity of the ten speakers was experimentally manipulated by teaching listeners to identify by name the talkers producing the natural speech utterances using a feedback-driven supervised training procedure. The sentences were sinewave replicas of the natural utterances. Subjects were trained (with feedback) until they were able to identify the ten talkers from the natural speech samples with at least 70% accuracy. After the training phase, subjects completed a sinewave transcription task designed to accustom them to the unusual acoustic qualities of sinewave replicas.

We compared two transcription presentation methods that differed in the amount of trial-by-trial variation among the sentences. For half the subjects, three repetitions of each sentence were presented randomly in the list, and subjects transcribed the entire sentence after each trial ("random presentation" condition). For the other half of the subjects, five repetitions of each sentence was presented consecutively in a blocked fashion ("blocked presentation" condition). Subjects were required to transcribe the entire sentence after the last repetition, although they could note partial information during any repetition. We assumed the latter method would facilitate transcription performance because there would be fewer trial-to-trial changes in the items, and more opportunities to attend to the sentence.

Voice recognition was then assessed using two generalization tasks in which listeners heard a new set of sentences and were required to identify the speaker. In one generalization test, the sentences were sinewave replicas whereas in the other generalization test, the sentences were naturally produced utterances. In both cases, the generalization tests used utterances that the subjects had not heard before during training. No feedback was provided. If prior exposure to sinewave speech is an important determinant of performance on the sinewave generalization test, there should be a gain in accuracy relative to the previous condition which did not include a transcription task (Experiment 2 from Sheffert et al., 1996).

Experiment 1

Method

Subjects

Twenty adult subjects were recruited from the Bloomington community. Of these, three subjects failed to complete the study due to work or school commitments, and one was excused because of a possible hearing impairment. The remaining sixteen subjects completed the natural speech training phase, the sinewave transcription task and the two generalization tests. All subjects were native speakers of American English and reported no history of a speech or hearing disorder at the time of testing. Subjects were paid for their participation.

Test Materials

Three types of sentences were used in the present experiment. The natural speech sinewave test sentences were the same items used by Sheffert at al. (1996), and developed by Remez et al. (1997). One set of sentences consisted of nine natural utterances produced by five male and five female talkers. Each talker produced all nine sentences, bringing the total number of sentences to 90. Audio recordings were obtained by asking speakers to read the sentences aloud in their natural speaking style. The sentences were recorded on audiotape in a sound-proof booth and were low-pass filtered at 4.5 kHz, digitally sampled at 10 kHz, equated for root mean squared (RMS) amplitude and stored as sampled data with 12-bit resolution.

The second set of sentences were sinewave replicas of the original natural speech tokens. To create these items, the frequencies and amplitudes of the first three formants were derived at 5 msec intervals. Formant values were obtained using two measures interactively: 1) linear predictive coding (LPC), and 2) discrete fourier transforms (DFT). Three time-varying sinusoids were then synthesized based on the center frequencies and amplitudes of the formants (Rubin, 1980). The synthesis algorithm preserved higher-order patterns of spectro-temporal change of the vocal tract transfer function, while eliminating the fundamental frequency, harmonic relations and fine-grained spectral information.

Three sentences were randomly selected (without replacement) for each of the three phases of the experiment (training, natural speech generalization and sinewave speech generalization). All sentences were rotated through all conditions for each listener to ensure that the observed effects were not due to any specific subset of the sentences or any order effects.

The third set of sentences were sinewave tokens used for the transcription test. These items consisted of 29 tonal analogs derived from natural speech utterances produced by a male speaker who was not part of the training set. The sentences were recorded, digitized and stored in the same manner as the stimulus materials described above.

Procedure

Training Phase

Listeners were trained to identify the names of the 10 speakers using the natural speech utterances. Subject testing was conducted for groups of three or less in a quiet listening room. During each training session, subjects heard a random ordering of five repetitions of three sentences from each of the 10 talkers (150 items total). The same three sentences were used for each talker in each training session, and subjects were told before hand which three sentences they would be hearing in a given session. The natural speech training sentences were presented binaurally to subjects at 75 dB SPL over matched and calibrated stereophonic headphones (Beyerdynamic DT100). Subjects were asked to listen carefully to each sentence and to pay close attention to the talkers' voice. Each time a sentence was presented, the subject was required to press one of ten keyboard buttons, each of which was labeled with a speaker's name. Keys 1-5 were labeled with female names and keys 6-10 with male names. Each time a subject made a response, the accuracy of that response and the name of the correct talker was displayed on the computer screen in front of the subject and recorded by the computer. Each training session lasted approximately 30 minutes. Training was continued until subjects achieved an average of 70% correct speaker recognition performance at the end of a session.

Sinewave Transcription Task

In this phase of the experiment, subjects completed the sinewave transcription task under the same listening conditions as the training task. The transcription task presented 29 sinewave sentences based on the natural speech production of a talker (R.E.R.) who was not part of our training ensemble. Subjects were told that the purpose of the task was to measure their ability to identify words in synthesized sentences. We encouraged the participants to listen in their "speech mode" by telling them that although the sentences sound very unnatural, they were in fact real speech that had been processed by a computer. Subjects were told that there was only one "voice" speaking the sentences. The listeners were instructed to transcribe the sentences as accurately as possible on a response form. They were not given feedback on their performance in this phase of the experiment.

Listeners in the "random presentation" group were told that each sentence would be repeated three times throughout the list. They were instructed to write down as many words or partial words as they could after each sentence, and guess if needed. The task was self-paced in order to give the listeners plenty of time to process and transcribe the sentences.

Subjects in the "blocked presentation" group were given the same instructions, with the exception that they were told that each sentence would be repeated five times in a row. Although subjects were required to write down the entire sentence after the last repetition, they were allowed to write down words or parts of words at anytime during the five repetitions. Listeners were given as much time as needed to listen to and write down the sentences. The transcription tasks took 20 to 30 minutes.

Familiarization Phase

Each of the generalization tests was preceded with a brief familiarization task designed to remind subjects of the correspondence between the training tokens and the names of the speakers. The task was simply an abbreviated version of a training session in which subjects listened and responded to one instance of each training sentence from each talker (30 items total). As in the training phase, the items were presented in a random order and subjects received feedback after each response. The familiarization task took approximately 8 minutes.

Generalization Tests

After reaching a 70% correct criterion in the natural speech training phase, subjects completed two generalization tests. One generalization test presented three novel sinewave sentences, whereas the second test presented three novel naturally produced sentences. All of the sentences presented during the generalization tests were new to the subjects. Half the subjects received the natural generalization test before the sinewave generalization test, whereas the other half received the tests in the opposite order. Each generalization test presented five repetitions of each of the three sentences from each talker in a random order (150 items total). Subjects were informed of the sentences they would be hearing before the start of each test. Subjects were asked to attend specifically to the talker's voice and to identify the talker by pressing one of the ten buttons on the keyboard as they had done in the previous training phase. Subjects did not receive feedback during either of the two generalization tests. Each generalization test lasted approximately 30 minutes.

Results

Training Performance

Examination of the training data revealed that listeners had little difficulty learning to identify the speakers from the natural sentences. All listeners reached criteria by the end of the third session. Seven subjects reached criterion after only a single training session. The mean number of training days was 1.75. Speaker identification performance on the last day of training averaged 79% for the subjects in the random presentation group, and 80% for subjects in the blocked presentation group. These values do not differ statistically from one another nor do they differ from the training data reported in Sheffert et al. (1996). The training data show that talker-specific aspects of the speech signal are easily attended to and that voices can be learned easily from sentence length materials.

Since there were no differences in the training performance across the two groups of subjects, all subsequent analyses of the training data were conducted on the combined scores from both groups of participants (N=16). Figure 1 displays identification performance on the last day of training as a function of talker. Overall, female talkers were identified better than male talkers (89% vs. 70% correct for female and male speakers, respectively). An ANOVA comparing talker identification performance on the last day of training revealed a significant effect of speaker sex, F(1, 158) = 46.98, p < .0001, confirming that the female speakers were more accurately identified than the male speakers.

Insert Figure 1 about here

The training data also revealed variability in the identifiability of different speakers within each sex. An ANOVA was conducted on the training scores for each sex. Reliable differences were found among the female speakers, F(4, 60) = 4.44, p < .003, and the male speakers, F(4, 60) = 7.82, p < .0001. Taken together, the natural speech training data show variability in the ease with which certain voices are identified. The patterning of the data across talkers is almost identical to the previous results of Sheffert et al. (1996).

Sinewave Transcription Performance

Performance on the sinewave transcription test was based on the mean number of syllables correctly identified in each sentence. Mean transcription performance for subjects in random presentation group was 49% compared to 42% for subjects in the blocked presentation group. This difference was not statistically significant.

To assess the effects of the different transcription methods on speaker recognition, we compared the data from the present experiment with the data from Experiment 2 in Sheffert et al. (1996). This earlier experiment did not include a sinewave transcription task. Table 1 displays the data from these three conditions. The table shows that the data from present experiment replicates the findings of Sheffert et al. Positive transfer was observed to the novel natural speech utterances, whereas very little transfer was observed to novel sinewave utterances. In addition, there was no benefit from the transcription task.

Table 1.

Proportion correct transcription and speaker recognition performance as a function of transcription presentation method.

	Transcription Presentation Method		
	Random	**Blocked**	**None***
Transcription	.49	.42	--
Training	.79	.80	.78
Natural Test	.82	.87	.88
Sinewave Test	.26	.28	.27

Note. * Data from Experiment 2 of Sheffert et al. (1996).

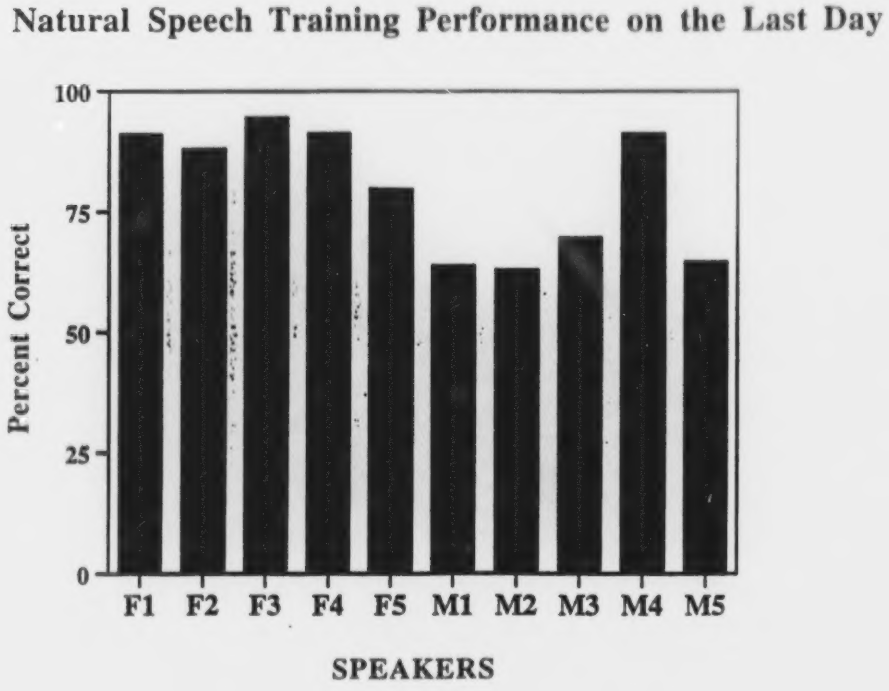


Figure 1. Mean speaker identification performance on natural speech for the last day of training as a function of speaker. F1 though F5 refer to the female speakers; M1 through M5 refer to the male speakers.

556

Separate ANOVA's were performed on the sinewave transcription scores, natural speech training performance, natural speech generalization and sinewave replica generalization. The statistical analysis confirms what is apparent in Table 1. For all measures, we found no differences in the mean test performance as a function of transcription method. Consequently, this variable was ignored in all further analyses and the data from the random and blocked transcription conditions were combined into a single group.

Generalization Performance

Because of differences in the identifiability of different speakers within the training set, the statistical analysis of the generalization tests was conducted using generalization scores to normalize for different levels of performance. The generalization scores were obtained by dividing the talker identification accuracy on the sinewave and natural speech generalization tests by talker identification accuracy on the training task.

At the time of the generalization testing, half of the subjects received the natural speech generalization test before the sinewave generalization test, whereas the other half completed the tests in the opposite order. Because we found no differences in the mean test performance as a function of test order, the two test order groups were pooled to form a single composite group. The statistical analysis of the generalization tests following training on natural speech was conducted using generalization scores.

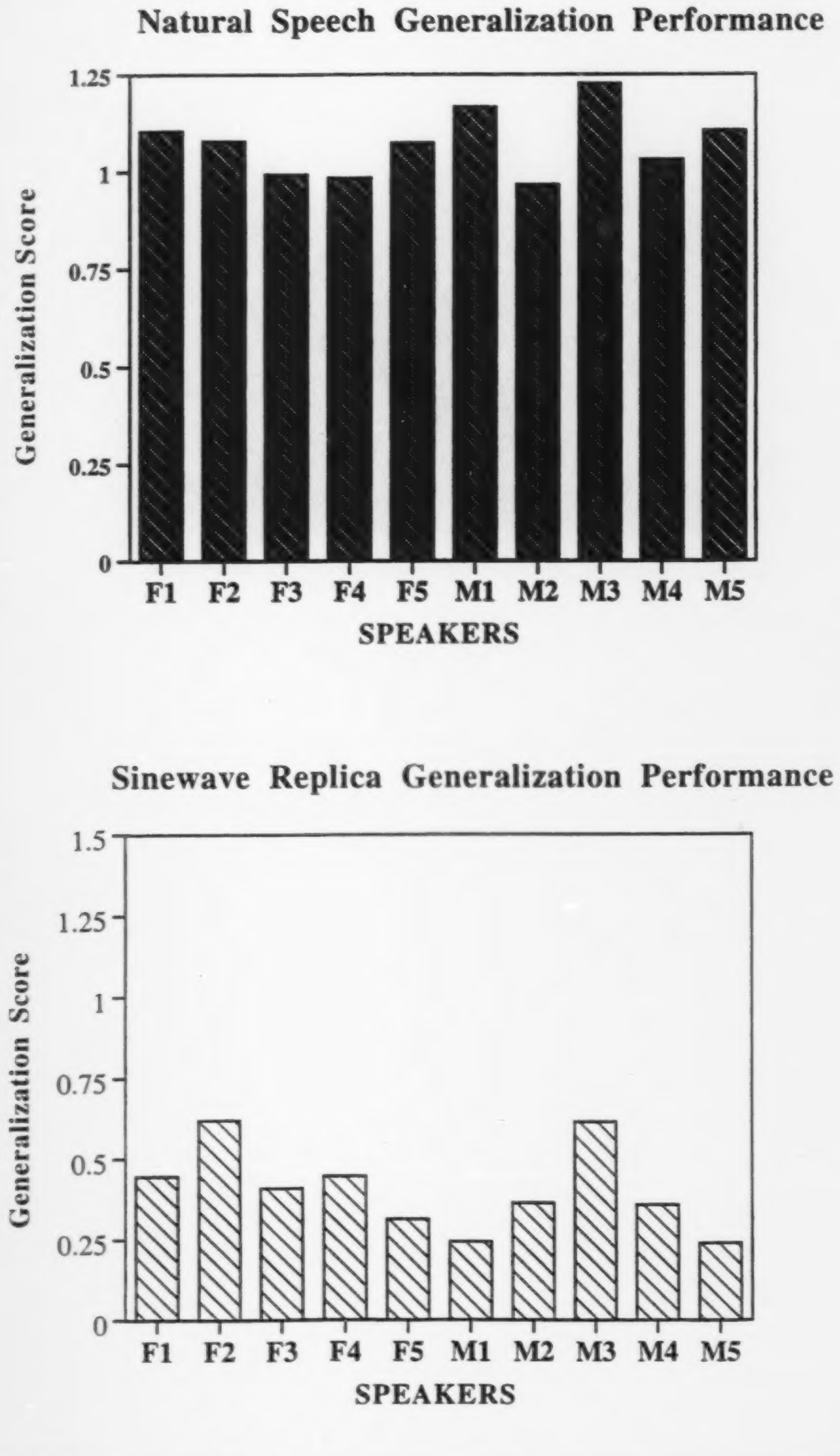
Figure 2 displays the generalization scores for the natural speech and sinewave replica generalization tests for each speaker. As in the previous perceptual training study, performance on the two generalization tests were very different. Specifically, speaker-specific knowledge acquired during the perceptual learning phase generalized to novel natural speech sentences, but not to novel sinewave replicas. Listeners' ability to recognize individuals from natural speech samples was 85%, as compared to only 27% for the sinewave samples. An ANOVA comparing the overall means from each of the three conditions (training, natural and sinewave) revealed a significant effect, $F(2, 45) = 179.88$, $p < .0001$. Planned comparisons revealed that training performance was significantly higher than performance on the sinewave replica generalization test [$t(15) = 19.82$, $p < .0001$], but was not reliably different than performance on the natural speech test. In addition, the two generalization test differed reliably from each other [$t(15) = 24.91$, $p < .0001$].

Insert Figure 2 about here

Male and female voices were recognized equally in both generalization tests, as shown in Figure 3. An ANOVA with the factor of speaker sex was conducted separately on the natural speech generalization scores and the sinewave replica generalization scores. The effect of speaker was not significant in either of the generalization tests.

Insert Figure 3 about here

Similarly, there were also no differences in the identifiability of voices among the female and male speakers in either generalization test condition, as shown in Figure 8. An ANOVA with the factor speaker


Natural Speech Generalization Performance

Sinewave Replica Generalization Performance

Figure 2. Mean speaker identification performance on the natural speech generalization (top panel) and sinewave replica generalization (bottom panel) as a function of training days and speaker sex. F1 through F5 refer to the female speakers; M1 through M5 refer to the male speakers.

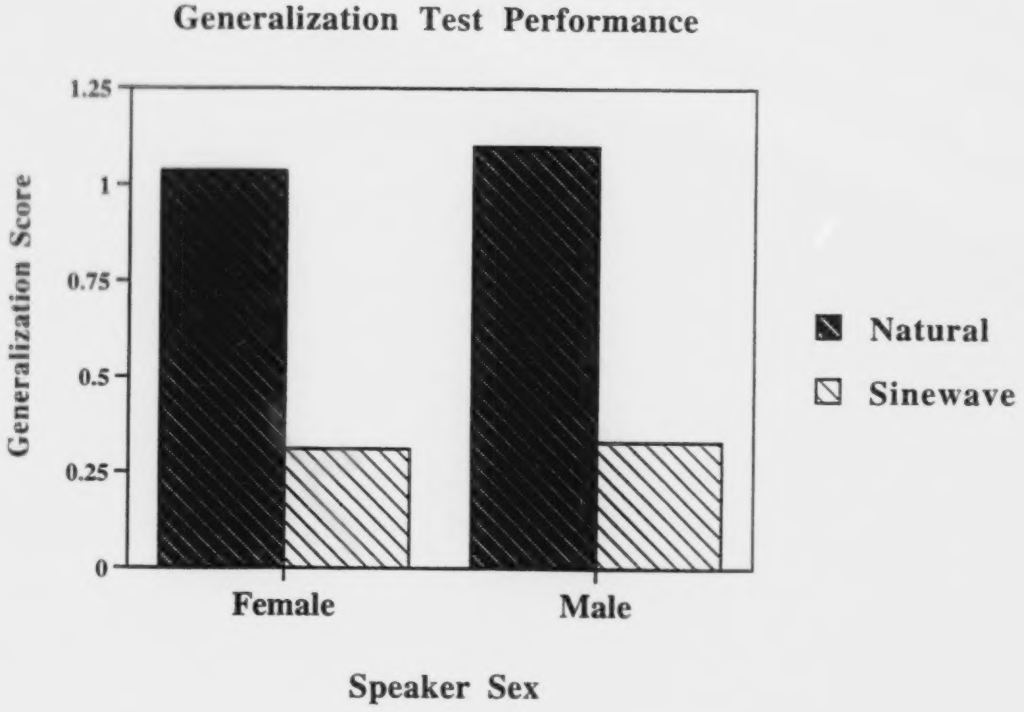


Figure 3. Mean generalization scores on the natural speech and sinewave replica generalization tests as a function of speaker sex.

was conducted on the natural speech generalization scores for each sex. The effect of speaker did not approach significance for either the female or the male speakers. No significant differences were found among the male speakers. For the sinewave replica generalization task, the effect of speaker was significant both for the female speakers $F(4, 60) = 4.66$, $p < .002$, and for the male speakers $F(4, 60) = 4.18$, $p < .005$.

The relationship between perceptual learning and generalization performance was assessed by comparing the relative ranking of the speakers across each condition. A Spearman's rho correlation was conducted between the training and generalization conditions. The analysis indicated that the speakers that were most easily identified during training were also the same speakers most easily identified in the natural test conditions (Spearman's rho = .891). The identifiability of talkers was less similar across the two generalization tests, as indicated by the modest correlation between the two condition (Spearman's rho = .552). The identifiability of talkers was least similar between training and the sinewave test (Spearman's rho = .345). This is another indication that the talker-specific knowledge acquired at training had little impact on performance in the sinewave generalization test. Note, also, that this correlation is somewhat lower than correlation of .515 reported in Sheffert et al. Although one must be cautious comparing measures across experiments, the lower correlation in the present study is consistent with the idea that familiarizing listeners with sinewave does not increase the relationship between training and sinewave test performance.

Discussion

The most important finding from the present study is that prior exposure and familiarity with sinusoidal speech did not improve subjects' ability to utilize talker-specific information contained in sinewave signals. The transcription task presented numerous samples of sinewave speech, with the expectation that subjects would become accustomed to the unusual sound of sinewaves and would become attuned to the phonetic properties of the signals, which would then facilitate sinewave voice recognition. This outcome did not occur. Generalization performance in this experiment was statistically indistinguishable from generalization performance from Sheffert et al. (1996), indicating that experience with sinewaves does not improve sinewave voice recognition.

It is possible that our subjects needed more substantial phonetic experience with sinewave utterances than our transcription task provided. Given that transcription performance was not particularly high (46% overall correct word identification), it may be that transcription was only beneficial to those subjects who could reliably extract phonetic information from the sinewaves. That is, listeners who had difficulty hearing the phonetic information contained in the sinewave replicas may also have had difficulty perceiving the voice information in the signals. If this is so, one would expect transcription performance to be related to sinewave voice recognition performance.

Examination of the individual subject data argues against this possibility. One example is a subject who was a good sinewave transcriber (56% correct word identification), yet had one of the lowest voice recognition scores on the sinewave generalization test (17% correct voice recognition). In general, the best transcribers (i.e., above 70% correct) were no more likely to outperform the worst transcribers (i.e., below 30% correct) on the sinewave generalization test. Thus, the degree to which a subject was able to perceive sinewaves as speech proved to be a poor predictor of their ability to extract talker-specific phonetic information and recognize voices from these patterns.

It is also possible is that the null effect of sinewave transcription occurred because the requirements of the task (word identification) did not overlap sufficiently with the requirements of the generalization test (voice recognition). If generalization performance is determined, in part, by the extent to which retrieval conditions match learning conditions, than perhaps subjects need to become familiar with the acoustic media during the course of voice learning, rather than word identification.

An alternative to the hypothesis tested in this report is that the poor voice transfer from natural speech to sinewave replicas may be due to differences in attention. Listeners may have focused on talker-specific information that was not present in the sinewave signals. Specifically, during natural speech learning, subjects may have focused their attention on the most obvious distinctive properties of speech that cue speaker identity - pitch, timbre or other suprasegmental characteristics. This information was absent from the sinewave test sentences. Poor generalization performance would be expected if listeners learned to distinguish talkers along perceptual dimensions in the natural speech that did not map onto the perceptual dimensions preserved in the sinewave signals.

In contrast, listeners trained from sinewave utterances were forced to rely on phonetic or segmental information to distinguish talkers, since other cues to talker identity were obliterated during sinusoidal synthesis. To a lesser extent, the same may be true of the listeners in Remez et al. (1997). Given that all these listeners were trained speech scientists or linguists and used to listening "phonetically", they may have focused more attention towards the properties of the natural speech that sinewaves possess. Consequently, they were often able to recognize their colleagues from a sinewave replica.

One way to determine if the pattern of results from the present experiment as well as those obtained in Experiment 2 of Sheffert et al. (1996) were due to differences in attention would be to assess transfer to a media that preserves many of the same suprasegmental cues to voice identity as natural speech while also being acoustically unusual and impoverished, like sinewave speech. Backwards speech fulfills these requirements. Backwards speech distorts temporally based fine-grained segmental information, such as information about consonants and diphthongs. Although some phonetic information is preserved, such as vowel quality, it is not enough to support word identification. As a result, backward speech is completely unintelligible. However, temporal reversal does not distort many suprasegmental aspects of speech. Characteristics based on long-term spectra such as fundamental frequency, F0 contour, speaking rate and formant relations are largely intact. Because these cues are important for speaker identification, it is relatively easy to recognize many familiar voices from reversed speech (Van Lancker, Kreiman, & Emmorey, 1995). If subjects are primarily exploiting suprasegmental information during natural speech training, than transfer from natural to reversed speech should be better than transfer to sinewaves. In addition, reversed speech sounds very strange. If transfer is positive, we will have further evidence that the peculiarity of the acoustic signal is not be the critical determinate of transfer in voice learning. This experiment, now currently underway in our lab, should provide additional insight into the learning and generalization of talker information.

References

Bricker, P.D., & Pruzansky, S. (1976). Speaker recognition. In N.J. Lass (Ed.), *Contemporary Issues in Experimental Phonetics* (pp. 295-326). New York: Academic Press.

Laver, J. (1991). *The gift of speech.* Edinburgh, Scotland: Edinburgh University Press.

Remez, R.E., Fellowes, J.M., & Rubin, P.E. (1996). Talker identification based on phonetic information. *Journal of Experimental Psychology: Human Perception and Performance*, **23**, 651-666.

BEST COPY AVAILABLE

Remez, R.E. Rubin, P.E., Pisoni, D.B., & Carroll, T.D. (1981). Speech perception without traditional speech cues. *Science*, **212**, 947-950.

Rubin, P.E. (1980) *Sinewave Synthesis.* Internal Memorandum, Haskins Laboratories, New Haven CT.

Sheffert, S. M., Pisoni, D. P., Remez, R. E. & Fellowes, J. M. (1996). Perceptual learning of natural and sinewave voices. In *Research on Spoken Language Processing Progress Report No. 20* (pp. 275-296). Bloomington, IN: Speech Research Laboratory, Indiana University

Van Lancker, D., Kreiman, J. & Emmorey, K. (1995). Familiar voice recognition: Patterns and parameters. Part I: Recognition of backward voices. *Journal of Phonetics,* **13**, 19-38.

Tongue Twisters Reveal Neighborhood Density Effects in Speech Production[1]

Michael S. Vitevitch

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405

---

[1] This research was supported by NIH-NIDCD Training Grant DC00012 to Indiana University.

M.S. Vitevitch

Tongue Twisters Reveal Neighborhood Density Effects in Speech Production

Abstract. The influence of similarity neighborhoods on speech production was investigated by experimentally eliciting speech errors with a tongue twister task. Half of the tongue twisters consisted of words from lexically dense neighborhoods, whereas the other half of the tongue twisters consisted of words from lexically sparse neighborhoods. The results showed that more erroneous repetitions were made on tongue twisters containing words from sparse neighborhoods than on tongue twisters containing words from dense neighborhoods. The implications of these findings for models of speech production and for models of word recognition are discussed.

A *phonological similarity neighborhood* refers to a group of similar sounding words that are confuseable with one another (Landauer & Streeter, 1973; Luce, 1986; Luce and Pisoni, 1998). One metric used to assess similarity is the addition, deletion, or substitution of a single phoneme (Greenberg & Jenkins, 1964). A number of studies have demonstrated that individual words differ in terms of neighborhood size, and that this difference has consequences on perception (Luce, 1986; Luce and Pisoni, 1998). That is, some words have many similar sounding words (i.e., a dense neighborhood), whereas other words have few similar sounding words (i.e., a sparse neighborhood). Words with dense neighborhoods are recognized more slowly and less accurately than words with sparse neighborhoods (Luce, 1986; Luce and Pisoni, 1998).

A number of other variables, including word frequency (Howes, 1957; Newbigginning, 1961; Savin, 1963; Solomon & Postman, 1952), phonotactic information (Vitevitch, 1997a; Vitevitch & Luce, in press; Vitevitch, Luce, Charles-Luce, & Kemmerer, 1997), and semantic information (Swinney, 1979), affect the speed and accuracy with which spoken words are recognized. Other studies have demonstrated that these variables also affect the speed and accuracy with which spoken words are *produced* (see Dell, 1988, 1990; Jescheniak & Levelt, 1994; Oldfield & Wingfield, 1965, for frequency effects; see Motley, 1973; and Motley & Baars, 1975, for phonotactic effects; see Motley and Baars, 1976a for a semantic effect). Little work, however, has investigated the influence of phonological similarity neighborhoods on speech production.

An earlier experiment by Goldinger and Summers (1989) demonstrated that neighborhood density influenced the voice onset time (VOT) for spoken words. They presented participants with word pairs of varying neighborhood density that differed in the voicing of the initial consonants (e.g., *dutch-touch*). They found that the differences in VOT between the first word and the second word of the pairs were larger for word pairs with dense neighborhoods than for word pairs with sparse neighborhoods. These differences became smaller for sparse neighborhood word pairs across sessions, but became larger for dense neighborhood word pairs across sessions.

Goldinger and Summers also found that the interword interval, or the time between the offset of the first word and the onset of the second word within each minimal pair, was greater for dense neighborhood word pairs than for sparse neighborhood word pairs. These results suggest that neighborhood density also affects the rate of speech production in demonstrable ways (see also Wright, 1998).

However, a regression analysis of reaction times from a picture naming task by Jescheniak and Levelt (1994) failed to find an effect of the number of words similar to a lexical item (as measured by the cohort count; Marslen-Wilson and Welsh, 1978; and as measured by the Coltheart-*N*; Coltheart, Davelaar, Jonasson, and Besner, 1977) on the rate of speech production. Because phonological similarity was not specifically manipulated in the Jescheniak and Levelt experiments, it is very likely that their posthoc analysis may have been insensitive to the influences of phonological similarity on speech production.

Although the results showing an influence of neighborhood density on the *speed* of speech production are not consistent, a study by Vitevitch (1997b) demonstrated a reliable influence of neighborhood density on the *accuracy* of speech production. Vitevitch analyzed the lexical characteristics of a corpus of malapropisms, or phonologically-based speech errors involving whole words (e.g., saying *monotony* instead of *monogamy*). The results from that investigation showed that malapropisms tend to have sparser neighborhoods than "control" words of equal length (in terms of number of phonemes) that were randomly selected from the lexicon. This corpus analysis suggests that neighborhood density has demonstrable affects on the *accuracy* with which a spoken word is produced.

However, several concerns have been raised regarding the use of data from speech error corpora. Specifically, Cutler (1982) and others (Ferber, 1991; MacKay, 1980; Mowrey and MacKay, 1990; Stemberger, 1992) have suggested that perceptual biases may distort what is perceived and recorded as a speech error, thereby disproportionately skewing the lexical characteristics of speech error corpora.

In order to avoid some of the potential sampling problems of speech error corpora, Stemberger (1992) has suggested that speech production findings be replicated either by examining multiple error corpora or by eliciting speech errors experimentally (see Baars, 1992; Motley & Baars, 1976b). Replication via experimentally eliciting speech errors offers several advantages over investigations of naturalistic error corpora.

Experimentally eliciting speech errors allows one to reduce the possibility of spurious effects resulting from perceptual biases by recording and carefully analyzing the experimental sessions. In addition, one can more precisely calculate actual error probabilities (Motley & Baars, 1976b), rather than estimate error probabilities from naturalistic error corpora. Finally, experimentally eliciting speech errors allows one to carefully examine and manipulate selected experimental variables.

To further investigate the influence of similarity neighborhoods on speech production, a tongue twister task—a task which elicits errors involving phonological segments rather than whole words (Baars, 1992)—was used to obtain speech errors on words of varying neighborhood density. Based on the findings involving a corpus of malapropisms (Vitevitch, 1997a), we would anticipate more errors involving phonological segments among tongue twisters containing words with sparse neighborhoods than among tongue twisters containing words with dense neighborhoods.

Because phonological errors occur at a different level of representation (i.e., a sub-lexical level) than whole-word malapropisms, which occur at a lexical or word-form level (Garrett, 1988; Levelt et al., 1991), it is also possible that a different pattern of errors may result. Thus, more errors involving phonological segments might occur among tongue twisters consisting of words with dense neighborhoods than among tongue twisters consisting of words with sparse neighborhoods. Because the goal of the present investigation is to examine empirically *if* similarity neighborhoods affect speech production, either outcome is acceptable.

Methods

Participants

Twenty-eight native English speakers from the Indiana University pool of Introductory Psychology students participated in partial fulfillment of a course requirement. All participants reported no history of a speech or hearing problem at the time of testing.

Stimuli

Ten pairs of highly confuseable target segments (those used in Experiment 2 of Shattuck-Hufnagel, 1992) were used to select CVC words for the tongue twisters used in this experiment. That is, only words that had the same initial segments as those in Experiment 2 of Shattuck-Hufnagel (1992) were considered in the selection of stimuli. (Two pairs of segments in Shattuck-Hufnagel (1992) were not used in the current experiment because a sufficient number of words with the desired characteristics could not be found in each condition.) Twenty tongue twisters, each containing four words, were created. Half of the tongue twisters were comprised of words with sparse neighborhoods, and the other half of the tongue twisters were comprised of words with dense neighborhoods.

Neighborhood density is defined as the number of words that are similar to the target item. Similarity is assessed by adding, deleting, or substituting one phoneme to the target word. The number of words found in an on-line version of the Webster's Pocket Dictionary, which contains nearly 20,000 words in a computer readable phonetic transcription, constituted the lexical neighborhood of a word.

Words from dense neighborhoods were items that had many similar sounding words found in the lexicon. The mean number of words in a neighborhood for items in the dense neighborhood condition was 23.9 words. Words from sparse neighborhoods were items that had few similar sounding words found in the lexicon. The mean number of words in a neighborhood for items in the sparse neighborhood condition was 15.4 words. The difference between the dense neighborhood and sparse neighborhood condition was highly significant ($F\,(1,78) = 143.20$, $p < .0001$).

Although the stimuli differed in neighborhood density, the words used in the sparse neighborhood tongue twisters and the dense neighborhood tongue twisters did not differ in log- frequency ($F\,(1,78) = 2.08$, $p = .15$) as measured by the Kucera & Francis (1967) word counts. The mean log-frequency of the items in the dense neighborhood tongue twisters was .83 occurrences per million, and the mean log-frequency of the items in the sparse neighborhood tongue twisters was 1.03 occurrences per million.

The two groups of words also did not differ in neighborhood log-frequency ($F\,(1,78) = 1.29$, $p = .25$). Neighborhood log-frequency is defined as the mean log-frequency of the words comprising the neighborhood (also assessed by counts from Kucera & Francis, 1967). The mean neighborhood log-frequency of the items in the dense neighborhood tongue twisters was 1.21 occurrences per million, and the mean log-frequency of the items in the sparse neighborhood tongue twisters was 1.15 occurrences per million. A complete listing of all the stimulus items is given in the appendix.

Finally, the words in each condition were also equivalent in familiarity ratings on a seven point scale. The scale ranged from (1) "don't know the word," to (4) "recognize the word, but don't know the meaning," to (7) "know the word" (Nusbaum, Pisoni, and Davis, 1984). The mean familiarity rating for items in the dense neighborhood tongue twisters was 6.69, and the mean familiarity rating for items in the sparse neighborhood tongue twisters was 6.80 ($F(1,78) = 1.20$, $p = .27$). Mean familiarity ratings above 5 ensured that almost all of the participants would be familiar with the words used in the tongue twisters.

Procedure

Participants were seated individually in a sound proof booth (IAC model 402) equipped with a CRT and a head-mounted microphone. The computer presented a prompt ("Please repeat the following words six times in a row.") and then randomly presented one tongue twister on the CRT for twelve seconds. Participants were instructed to repeat the tongue twister six times as *quickly* as they could. The tongue twister remained on the CRT for the entire duration. Responses were recorded on high quality audiotape for later analysis. At the end of twelve seconds, the prompt was flashed on the CRT and a new trial began.

A practice session of five pseudo-tongue twisters (i.e., four randomly selected words from the items that did not have an initial consonant which matched the Shattuck-Hufnagel (1992) highly confuseable segment pairs) were used to familiarize the participants with the task. The responses from these stimuli were not included in the final analyses.

Results

Participants were required to repeat each tongue twister six times. A repetition was scored as a speech error if a perseveration, anticipation, or exchange of the initial consonants was produced. Work by Shattuck-Hufnagel (1979; Shattuck-Hufnagel & Klatt, 1979) suggests that the first replacement in an exchange is the true cause of the exchange error; the second replacement occurs by default. Thus, to avoid inflating the error rate, individual phoneme errors were not counted and used in calculating the error rate. Rather, the entire repetition was scored as either correct or as an erroneous repetition if one or more errors occurred in that attempt to repeat the tongue twister.

Misreadings or insertions of segments that were not conditioned by the stimuli were not counted as speech errors because these errors presumably arise from sources other than the levels of representation under investigation. See Table 1 for examples of responses that were counted as errors and responses that, although not correct, were not counted as errors.

Table 1.

Examples for the tongue twister "dash gab gaze doubt."

Responses counted as errors	
Examples	*Reason response counted*
dash dab gaze doubt	perseveration
gash gab gaze doubt	anticipation
gash dab gaze doubt	exchange
Responses not counted as errors	
Examples	*Reason response not counted*
bash gab gaze doubt dash gab glaze doubt	misreading or unconditioned error

A one-way repeated measures ANOVA (dense vs. sparse neighborhood) was performed on the total number of erroneous repetitions for each condition of the 28 participants. A highly significant difference was observed (F_1 (1, 27) = 16.88, p = .0003; F_2 (1, 18) = 6.16, p = .02). More erroneous repetitions were found among tongue twisters containing words from sparse neighborhoods than among tongue twisters containing words from dense neighborhoods. The results are shown in Table 2.

Table 2.

Mean number of erroneous repetitions
(Standard deviations are in parenthesis)

	Neighborhood Density	
	Sparse	*Dense*
Number of Erroneous Repetitions	7.5 (5.4)	4.2 (3.0)

Discussion

The results of the present speech error elicitation experiment found that more erroneous repetitions were produced among tongue twisters comprised of words with sparse neighborhoods than among tongue twisters comprised of words with dense neighborhoods. These results demonstrate that neighborhood density does affect the *accuracy* with which words are produced.

The present findings also replicate one of the results obtained from a corpus analysis by Vitevitch (1997b). Specifically, words with sparse neighborhoods are more prone to mis-productions than words with more dense neighborhoods. The current results extend the findings of Vitevitch (1997b) by demonstrating these effects with an experimental task which elicits speech errors rather than relying on corpora analyses. Furthermore, these results demonstrate that phonological similarity neighborhoods also influence speech production at other—sub-lexical—levels not just the word-form level, as the analysis of whole-word speech errors known as malapropisms suggests.

The present findings also have several implications for speech production and language processing in general. The influence of neighborhood density on phonological speech errors suggests that the activation of similar lexical items affects the processing of sublexical items. The influence of neighboring word-forms on sublexical representations may be easily accounted for by an interactive model of speech production (Dell, 1986, 1988, 1990; Harley, 1984). Dell (1986, 1988, 1990), for example, claims that activation spreads in a bi-directional manner between word-form representations and sub-lexical representations, thereby strengthening the activation of the desired word-form. It is unclear how the present results could be accounted for in serial-processing models of speech production (Garrett, 1980, 1988; Levelt et al., 1991).

Within an interactive activation framework similar to Dell's (1986, 1988, 1990), it is hypothesized that neighboring word-forms also activate sublexical representations. (Activation of neighboring word-forms may occur via activation spreading laterally from the intended word-form, or via activation spreading up from the sublexical level to all the word-forms that contain that sublexical representation.)

The activation that spreads from neighboring word-forms to sublexical items further strengthens the sublexical representations that comprise the intended word and, therefore, the intended word-form itself.

Intended word-forms in dense neighborhoods would presumably receive supportive activation from many neighboring word-forms, whereas intended word-forms in sparse neighborhoods would receive supportive activation from few neighboring word-forms via the sublexical level. Such a model would predict that intended word-forms that receive larger amounts of supportive activation (i.e., those in dense neighborhoods) would be less prone to phonological errors. Furthermore, intended word-forms that receive smaller amounts of supportive activation (i.e., those in sparse neighborhoods) would be more prone to phonological errors (see Taraban & McClelland, 1987, for a discussion of "gang effects"). How a serial-processing model of speech production (Garrett, 1980, 1988; Levelt et al., 1991) would account for the influence of neighborhood density on phonological speech errors is unclear.

The influence of variables such as word frequency, phonotactic information, and (as illustrated in the current study) neighborhood density on both speech production and speech perception raises some interesting questions about the processes and representations that are used in speech production and speech perception. Do speech production and speech perception access separate sets of representations? Alternatively, are there a common set of representations that are affected by the same lexical variables in different ways (depending on which "direction"—phonemes-to-concepts or concepts-to-phonemes—processing proceeds)? A related issue is whether models which have been proposed for only speech production or only speech perception can also account for the other process.

Although an early account of word recognition (Forster, 1978) also attempted to explain the production of words (and the mis-production of words), ensuing models proceeded to account for *either* perception (Luce & Pisoni, 1998; Marslen-Wilson and Welsh, 1978; McClelland & Elman, 1986; Norris, 1994) *or* production (Dell, 1986, 1988, 1990; Garrett, 1980, 1988; Harley, 1984; Levelt et al., 1991). In explicitly modeling only one process, researchers may have implicitly suggested that the representations and processes involved in speech production and speech perception may be separable.

A number of researchers, however, have suggested that production and perception are intimately interconnected (see, for example, Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967; O'Seaghdha & Marin, 1997; Pisoni, Svirsky, Kirk, & Miyamoto, 1997). In light of this issue, it would be interesting to see if the neighborhood activation model (a model of spoken word recognition that was designed to account for neighborhood density effects in word recognition) can account for the current findings. The neighborhood activation model (NAM) proposes that similar sounding word-forms compete among each other in the word recognition process (Luce and Pisoni, 1998). If an incoming sound pattern activates many similar sounding items (i.e., a dense neighborhood) in the lexicon, word recognition will be slower and less accurate than for an incoming sound pattern which activates few similar sounding items (i.e., a sparse neighborhood) in the lexicon (Luce and Pisoni, 1998).

Because word frequency influences perception and production in an analogous manner (i.e., high frequency words are produced and recognized more quickly and more accurately than low frequency words), one might expect that neighborhood density affects production and perception analogously. That is, more speech errors should be found among words with dense neighborhoods. The results of the present experiment (see also Harley & Brown, in press; Vitevitch, 1997b) show that the opposite pattern is obtained in production; namely, more errors are found among words in sparse neighborhoods than dense neighborhoods.

The inability of NAM to account for the production findings could be due to a number of factors. One reason may be that NAM was designed as a model of word recognition not speech production. Despite many similarities between models of speech production and word recognition (e.g., in the representations they posit), it may be that no current model of word recognition can also account for speech production effects (and vice-versa). However, in light of the assumption that perception and production are interconnected (Liberman et al., 1967; O'Seaghdha & Marin, 1997; Pisoni et al., 1997) such an explanation is unpalatable; a model of language processes should be able to account for both speech production and speech perception.

Alternatively, the current architecture of NAM may be insufficient to account for both speech production and speech perception data. The present instantiation of NAM consists of only a word-form level in which competition among similar sounding items occurs. Perhaps a sublexical level must be added to NAM in order for it to fully account for speech production as well as word recognition effects. Indeed, such a proposal was made recently by Vitevitch (1997b) in an effort to account for the results of the malapropism error corpus, as well as by Vitevitch (1997a) and Vitevitch and Luce (in press) in order to account for effects of phonotactics on spoken word recognition.

Although the present results do not directly address the broader issue of how language production and comprehension are interfaced, they do demonstrate that neighborhood density affects the accuracy with which words are produced. Specifically, words in dense neighborhoods tend to be produced more accurately than words in sparse neighborhoods. These findings are the opposite of neighborhood density effects in word recognition. (Recall that in word recognition, words in dense neighborhoods are recognized less accurately than words in sparse neighborhoods.) The findings from the present experiment contribute to a growing body of literature (Harley & Brown, in press; Vitevitch, 1997b) suggesting that lexical neighborhoods influence speech production in addition to speech perception. Neighborhood density effects in speech production pose a serious challenge to all contemporary accounts of speech production based on serial-processing models (e.g. Garrett, 1980, 1988; Levelt et al., 1991).

References

Baars, B.J. (1992). A dozen competing-plans techniques for inducing predictable slips in speech and action. In B.J. Baars (Ed.), *Experimental slips and human error: Exploring the architecture of volition.* (pp. 129-150) Plenum Press: New York.

Coltheart, M., Davelaar, E., Jonasson, J.T. and Besner, D. (1977). Access to the internal lexicon. In S. Dornic (Ed.), *Attention and performance VI* (pp. 535-555). Hillsdale, NJ: Erlbaum.

Cutler, A. (1982). The reliability of speech error data. In A. Cutler (Ed.), *Slips of the tongue and language production* (pp. 7-28). Berlin: Walter de Gruyter/Mouton; also appeared in *Linguistics* (1981) **19**, 561-582.

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review*, **93**, 283-321.

Dell, G. S. (1988). The retrieval of phonological forms in production: Tests of predictions from a connectionist model. *Journal of Memory and Language*, **27**, 124-142.

Dell, G. S. (1990). Effects of frequency and vocabulary type on phonological speech errors. *Language and Cognitive Processes*, **5**, 313-349.

Ferber, R. (1991). Slip of the tongue or slip of the ear? On the perception and transcription of naturalistic slips of the tongue. *Journal of Psycholinguistic Research*, **20**, 105-122.

Forster, K.I. (1978). Accessing the mental lexicon. In E. Walker (Ed.), *Explorations in the biology of language*. (pp. 139-174). Montgomery, VT: Bradford.

Garrett, M.F. (1980). Levels of processing in sentence production. In B. Butterworth (Ed.), *Language Production, Volume I: Speech and talk*. (pp. 177-220). London: Academic Press.

Garrett, M.F. (1988). Processes in language production. In F.J. Newmeyer (Ed.), *Linguistics: The Cambridge survey: Vol III Language: Psychological and biological aspects* (pp. 69-96). Cambridge, UK: Cambridge University Press.

Goldinger, S.D. and Summers, V.W. (1989). *Lexical neighborhoods in speech production: A first report*. Research on Speech Production (Progress Report No. 15) Bloomington: Indiana University, Department of Psychology.

Greenberg, J. H., and Jenkins, J. J. (1964) Studies in the psychological correlates of the sound system of American English. *Word*, **20**, 157-177.

Harley, T.A. (1984). A critique of top-down independent level models of speech production: Evidence from nonplan-internal speech errors. *Cognitive Science*, **8**, 191-219.

Harley, T.A. and Brown, H.E. (in press). What causes a tip-of-the-tongue state? Evidence for lexical neighbourhood effects in speech production. *British Journal of Psychology*.

Howes, D.H. (1957). On the relation between the intelligibility and frequency of occurrence of English words. *Journal of the Acoustical Society of America*, **29**, 296-305.

Jescheniak, J.D. and Levelt, W.J.M. (1994). Word frequency effects in speech production: Retrieval of syntactic information and of phonological form. *Journal of Experimental Psychology: Learning, Memory and Cognition*, **20**, 824-843.

Kucera, H. and Francis, W.N. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.

Landauer, T. K. and Streeter, L.A. (1973). Structural differences between common and rare words: Failure of equivalence and assumptions for theories of word recognition. *Journal of Verbal Learning and Verbal Behavior*, **12**, 119-131.

Liberman, A.M., Cooper, F.S., Shankweiler, D.P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, **74**, 431-461.

Levelt, W.J.M., Schriefers, H., Vorberg, D., Meyer, A.S., Pechmann, T.,and Havinga, J. (1991). the time course of lexical access in speech production: A study of picture naming. *Psychological Review*, **98**, 122-142.

Luce, P.A. (1986). *Neighborhoods of words in the mental lexicon*. Doctoral dissertation, Indiana University, Bloomington, IN.

Luce, P.A. and Pisoni, D.B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing*, **19**, 1-36.

MacKay, D.G. (1980). Speech errors: Retrospect and prospect. In V.A. Fromkin (Ed.), *Errors in Linguistic Performance: Slips of the tongue, ear, pen, and hand* (pp. 319-332). New York: Academic Press.

Marslen-Wilson, W. and Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology*, **10**, 29-63.

McClelland, J.L. and Elman, J.L. (1986). The Trace model of speech perception. *Cognitive Psychology*, **18**, 1-86.

Motley, M.T. (1973). An analysis of spoonerisms as psycholinguistic phenomena. *Speech Monographs*, **40**, 66-71.

Motley, M.T., and Baars, B.J. (1975). Encoding sensitivities to phonological markedness and transitional probability : Evidence from spoonerisms. *Human Communication Research*, **2**, 351-361.

Motley, M.T., and Baars, B.J. (1976a). Semantic bias effects on the outcome of verbal slips. *Cognition*, **4**, 177-187.

Motley, M. T. and Baars, B. J. (1976b). Laboratory induction of verbal slips: A new method for psycholinguistic research. *Communication Quarterly*, **24**, 28–34.

Mowrey, Richard A. and MacKay, Ian R. (1990). Phonological primitives: Electromyographic speech error evidence. *Journal of the Acoustical Society of America*, **88**, 1299-1312.

Newbigging, P.L. (1961). The perceptual reintegration of frequent and infrequent words. *Canadian Journal of Psychology*, **15**, 123-132.

Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. *Cognition*, **52**, 189-234.

Nusbaum, H.C., Pisoni, D.B. and Davis, C.K (1984). *Sizing up the Hoosier mental lexicon: Measuring the familiarity of 20,000 words*. Research on Speech Perception, Progress Report no. 10. Speech Research laboratory. Psychology Department, Indiana University, Bloomington, Indiana.

Oldfield, R.C. and Wingfield, A. (1965). Response latencies in naming objects. *Quarterly Journal of Experimental Psychology*, **17**, 273-281.

O'Seaghdha, P.G. and Marin, J.W. (1997). Mediated semantic-phonological priming: Calling distant relatives. *Journal of Memory and Language*, **36**, 226-252.

Pisoni, D.B., Svirsky, M.A., Kirk, K.I., and Miyamoto, R.T. (1997). Looking at the "stars": A first report on the intercorrelations among measures of speech perception, intelligibility, and language in pediatric cochlear implant users. Paper presented at the Vth International Cochlear Implant Conference, May 1-3, 1997, New York, NY.

Savin, H.B. (1963). Word-frequency effect and errors in the perception of speech. *Journal of the Acoustical Society of America*, **35**, 200-206.

Shattuck-Hufnagel, S. (1979). Speech errors as evidence for a serial order mechanism in sentence production. In W.E. Cooper & E.C.T Walker (Eds.), *Sentence Processing: Psycholinguistic studies presented to Merrill Garrett*. Hillsdale, NJ: Erlbaum.

Shattuck-Hufnagel, S. (1992). The role of word structure in segmental serial ordering. *Cognition*, **42**, 213-259.

Shattuck-Hufnagel, S. & Klatt, D. (1979). The limited use of distinctive features and markedness in speech production: Evidence from speech error data. *Journal of Verbal Learning and Verbal Behavior*, **18**, 41-55.

Solomon, R.L. and Postman, L. (1952). Frequency of usage as a determinant of recognition thresholds for words. *Journal of Experimental Psychology*, **43**, 195-201.

Stemberger, J.P. (1992). The reliability and replicability of naturalistic speech error data: A comparison with experimentally induced errors. In B.J. Baars (Ed.), *Experimental slips and human error: Exploring the architecture of volition*. (pp. 195-215) Plenum Press: New York.

Swinney, D. A. (1979). Lexical access during sentence comprehension: (Re)consideration of context effect. *Journal of Verbal Learning and Verbal Behavior*, **18**, 645-659.

Taraban, R. and McClelland, J.L. (1987). Conspiracy effects in word pronunciation. *Journal of Memory and Language*, **26**, 608-631.

Vitevitch, M.S. (1997a). *Phonotactics and spoken word recognition*. Unpublished doctoral dissertation, University at Buffalo, Buffalo, NY.

Vitevitch, M.S. (1997b). The neighborhood characteristics of malapropisms. *Language and Speech*, **40**, 211-228.

Vitevitch, M.S. and Luce, P.A. (in press). When words compete: Levels of processing in spoken word perception. *Psychological Science*.

Vitevitch, M.S., Luce, P.A., Charles-Luce, J., and Kemmerer, D. (1997). Phonotactics and syllable stress: Implications for the processing of spoken nonsense words. *Language and Speech*, **40**, 47-62.

Wright, R. A. (1998). Lexical sources of variation in production: Neighborhood effects. This volume.

Appendix

LOW DENSITY TONGUE TWISTERS	HIGH DENSITY TONGUE TWISTERS
balm peach pig bull	wail reek reel weed
rage weep wave rise	meal nail neat mole
map noose noon mead	fat pill pin fit
fig pawn pave fad	peep kit kin pick
page case cave palm	lace road rock lag
leaf rice robe lung	gill beer bun goal
birth gang gash bob	tuck den dial ton
deem tame tide dose	pad bile bout par
save shell shun sour	shack seep sip shock
dash gab gaze doubt	gore dame dill gull

Some Factors Affecting Recognition of Spoken Words by Normal Hearing Adults[1]

Ann R. Bradlow,[2] Gina M. Torretta,[3] and David B. Pisoni

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405

---

[1] This research was supported by NIH-NIDCD Training Grant DC-00012 and by NIH-NIDCD Research Grant DC-00111 to Indiana University. We are grateful to Luis Hernandez for technical support.

[2] Department of Communication Sciences and Disorders, Northwestern University, Evanston, IL

[3] Central Institute for the Deaf, St. Louis, MO

605

Some Factors Affecting Recognition of Spoken Words by Normal Hearing Adults

Abstract. An analysis of intelligibility data from a carefully constructed database of recorded speech was conducted in order to investigate the combined effects of various talker-, listener-, and item-related characteristics that contribute to variability in intelligibility of isolated words. Materials came from the Indiana Multi-Talker Word Database, which consists of a set of recorded words from multiple talkers at three speaking rates along with intelligibility data in the form of transcriptions by a large number of native English listeners. Results showed a strong effect of lexical discrimination ("easy" words had higher intelligibility scores than "hard" words), and a strong effect of speaking rate (slow and medium rate words had higher intelligibility scores than fast rate words). Furthermore, we observed a complex relationship between the various factors whereby the difficulties imposed by one factor, such as a fast speaking rate or an inherently difficult lexical item, could be overcome by the advantage gained by the listener's experience with the speech of a particular talker. Implications of these findings for the development of spoken language processing assessment instruments and assistive devices for "special listener populations" are discussed.

Introduction

It is well known that even under ideal listening and speaking conditions, transmission accuracy of the speaker's intended message to the listener often varies greatly. Recent work in our laboratory has focused on some of the factors that contribute to the observed variability in normal speech intelligibility. To date, several factors have been shown to directly influence overall speech intelligibility. First, the degree of variability in the stimulus materials has been show to have a major impact on the listener's speech recognition accuracy. For example, word recognition accuracies decrease and response times increase when listeners are presented with spoken word lists that incorporate a high-degree of stimulus variability due to the presence of multiple talkers and speaking rates, relative to spoken word lists in which such stimulus variability is minimized (Mullennix et al., 1989; Sommers et al, 1994). Second, familiarity on the part of the listener's with the talker's voice and articulatory characteristics enhances word recognition accuracy under difficult listening conditions. For example, Nygaard, Sommers and Pisoni (1994) showed that listeners were more accurate at identifying words in noise when spoken by a familiar talker than when spoken by a novel talker. Third, lexical characteristics of the particular words in a stimulus set exert a strong influence on overall intelligibility. Several studies have shown that "easy" words (i.e., words with few phonetically similar "neighbors" with which they could be confused) have a distinct intelligibility advantage over "hard" words (i.e., highly confusable words with many phonetically similar neighbors) (Pisoni et al., 1985; Luce, 1986; Luce et al., 1990). Finally, in a recent study of the talker-specific acoustic-phonetic characteristics that correlate with inter-talker intelligibility differences, Bradlow et al. (1997) showed that talkers who exhibited a high-degree of "articulatory precision" in their speech generally had higher overall speech intelligibility scores than talkers who tended to produce more "reduced" speech. Taken together, these studies demonstrate the range of stimulus-, listener- and talker-related factors that combine to result in the observed variability in normal speech intelligibility.

The present study continues this line of research by investigating the combined effects of various talker-, listener-, and item-related characteristics that contribute to overall intelligibility of isolated words in a carefully constructed database of recorded speech. Materials for this study came from the Indiana Multi-Talker Word Database (Torretta, 1995), which consists of a set of recorded words from multiple talkers at three speaking rates along with intelligibility data in the form of transcriptions by a large number of native English listeners. This intelligibility data provided us with the means to assess the combined effects of speaking rate, lexical discrimination, and listener-talker adaptation on isolated word intelligibility by native listeners. By directly examining the effects of these characteristics on native-language word intelligibility, we hoped to obtain a baseline measure of normal variability in isolated word recognition that could then serve as a basis for comparison of word recognition performance by a variety of listeners under various presentation conditions. For example, non-native and hearing-impaired listeners appear to be particularly sensitive to stimulus variability and adverse listening conditions, such as in the presence of multiple-talkers or background noise. Thus, knowledge about the factors that affect normal speech intelligibility by normal listeners may be particularly useful for the development of spoken language processing assessment instruments and assistive devices for "special listener populations."

Method

The "Easy" and "Hard" Word Lists

An "easy" list and a "hard" list of words (75 items each) were compiled such that the two lists differed in terms of three lexical characteristics (Pisoni et al., 1985; Luce, 1986; Luce et al., 1990; Luce and Pisoni, 1997). First, using the word frequency counts provided by the Brown Corpus of printed text (Kucera and Frances, 1967), the words were selected such that the mean word frequency of the easy list was substantially higher than that of the hard list (309.7 vs. 12.2 per million). Second, using an on-line version of Webster's Pocket Dictionary (20,000 entries) in conjunction with a custom-designed lexical search program, words were selected such that the neighborhood density (the number of phonetic "neighbors") of the easy list was lower than that of the hard list (13.5 vs. 26.6). In these neighborhood density counts, a neighbor of a given word was defined as any word that differed from the target word by a one phoneme addition, substitution or deletion in any position. Third, the two word lists were constructed such that the neighborhood frequency (the mean frequency of the neighbors) of the easy list was much lower than that of the hard list (38.3 vs. 282.2 per million). The net result of these three word selection criteria, was that the "easy" list consisted of a set of words that are frequent in the language, and that have few phonetically-similar, low-frequency neighbors with which they could be confused. In contrast, the "hard" list consisted of words with many neighbors that are high in frequency relative to the target word. Easy words "stick out" from sparse neighborhoods; hard words are "swamped" by dense neighborhoods. Finally, in order to ensure that subjects would be familiar with all of the words in both lists, all words were judged as highly familiar by normal-hearing adults, i.e., received a familiarity rating of 6.7 or higher on a 7 point scale where 1 indicated the lowest and 7 indicated the highest degree of familiarity (Nusbaum et al., 1984).

Digital Speech Recordings

Ten talkers (five males and five females) were recorded producing both the easy and the hard word lists at three different speaking rates (fast, medium, and slow), giving a total of 4500 tokens (150 words x 3 speaking rates x 10 talkers). None of the talkers had any known speech or hearing impairments at the time of recording, and all were native speakers of General American English. The talkers were told in advance that they would be asked to produce three word lists of 150 words each at three different speaking rates. Each individual talker was allowed to regulate his/her own speaking rate, so long as the three rates were distinct. An analysis of the word durations for each talker at each of the three rates, confirmed that each talker successfully produced the three lists with three distinct speaking rates. The mean durations were 809 ms (range 576-1030 ms), 525 ms (range 466-579 ms), and 328 ms (range 264-413 ms) for the slow, medium, and fast words, respectively.

All 150 words (75 easy plus 75 hard) were presented to the talkers in random order on a CRT monitor in a sound-attenuated booth (IAC 401A). The stimuli were transduced with a Shure (SM98) microphone, and digitized on-line (16-bit analog-to-digital converter (DSC Model 240) at a 20 kHz sampling rate). The recordings were all live-monitored by an experimenter for gross misarticulations and hesitations. Each individual digital file was then edited by hand to remove the silent portions at the beginning and end of each word file. The average root means square amplitude of each of the digital speech files was then equated. Finally, the files were converted to PC WAV format for presentation to listeners using a PC-based perceptual testing system (Hernandez, 1995).

Speech Intelligibility Tests

Speech intelligibility scores were collected from independent groups of ten normal-hearing listeners, each of whom transcribed the full set of 150 words from one talker at one speaking rate, for a total of thirty groups of ten listeners (10 talkers x 3 speaking rates). The words were presented to the listeners in random order over matched and calibrated DT-100 headphones via a PC-based perceptual testing system (Hernandez, 1995). The words were presented in the clear (no background noise was added) at a comfortable listening level (75 dB/SPL). On each trial, the listeners heard the word and then typed in the response on the computer keyboard. In the data scoring, a word was counted as correct if all of the letters were present and in the correct order, if all the letters were present but not in the correct order, or if the transcribed word was a homophone of the intended word.

These transcription scores provided a means of investigating the effects of speaking rate (fast vs. medium vs. slow) and lexical discrimination (easy vs. hard) on isolated word intelligibility. Additionally, since each group of listeners transcribed the full set of 150 words by a single talker at a single rate in a single transcription session, we could also use these intelligibility data to investigate whether listeners adapted to talker-specific characteristics to the extent that the intelligibility scores improved from the beginning to the end of the transcription session. We hypothesized that this kind of listener-talker "attunement" on the part of the listener, which occurs over the course of exposure to the speech of a particular talker, would interact with the lexical (easy vs. hard) and speaking-rate (fast vs. medium vs. slow) factors such that there would be a greater listener-talker adaptation effect as the other factors increased in difficulty. Such a finding would indicate that listener-talker familiarity can compensate for the word recognition difficulties associated with increased speaking rate and easily-confused lexical items.

Results

Figure 1 shows the overall percent correct transcription scores across all talkers and listeners for the easy and hard word lists at each of the three speaking rates. As expected based on earlier investigations of the effects of these lexical characteristics on speech perception (Pisoni et al., 1985; Luce, 1986; Luce et al., 1990), the easy word lists were generally more accurately transcribed than the hard word lists. As shown in Table I, the higher transcription accuracy for the easy list relative to the hard list held true for almost all speakers at all three speaking rates. The exception were for Talkers 1, 5, 6 and 9 at the slow rate, where there was no easy-hard difference, and for Talker 6 at the medium rate where there was a very small advantage for the hard word list. Thus, the word identification advantage for easy words over hard words is a highly robust effect that generalizes across multiple talkers and speaking rates.

Insert Figure 1 about here

Figure 1 also shows a substantial decline in transcription accuracy for the fast rate relative to the medium and slow rates for both the easy and the hard word lists. However, there was little difference in transcription accuracy between the slow and medium rate words. This pattern of results was somewhat surprising in view of the fact that, on average, the slow words were about 54% longer than the medium words. Thus, it appears that isolated word intelligibility is not enhanced by slowing the speaking rate.

These findings were all confirmed by a repeated-measures ANOVA (nested design) with both rate (fast, medium, slow) and lexical category (easy, hard) as within subject variables, and the intelligibility scores for each talker in each condition averaged across all ten listeners as the dependent variable (see Table I). There was a main effect of rate (F(2,18)=7.456, p=.0013), and a main effect of lexical category (F(1,18)=20.111, p=.0015). An examination of the contrasts showed a significant difference (at the p<.005 level) between the fast and medium rates for both the easy and the hard words, but no difference between the medium and slow rates for either the easy or the hard words. Furthermore, at all three rates, the easy vs. hard difference was significant at the p<.005 level.

TABLE 1.

Mean intelligibility scores across all ten listeners for the easy and hard word lists by each talker at each speaking rate.

	Easy			Hard		
Talker	**Slow**	**Medium**	**Fast**	**Slow**	**Medium**	**Fast**
1	91.07	92.40	86.13	82.67	81.20	72.27
2	94.40	95.47	94.27	94.80	94.40	89.33
3	94.67	94.00	94.93	88.93	89.60	92.53
4	92.40	96.00	88.27	88.67	87.20	78.00
5	94.00	94.40	86.27	89.47	91.33	75.47
6	92.93	93.87	91.87	92.80	90.40	89.73
7	90.67	89.20	89.47	91.07	90.26	87.87
8	94.93	96.27	92.93	93.60	88.40	89.47
9	95.07	96.67	95.73	92.40	92.13	84.40
10	95.07	98.40	96.27	94.93	95.46	90.67
mean	93.52	94.67	91.61	90.93	90.04	84.97

The next step in our analysis of these intelligibility data was to investigate whether isolated word intelligibility can be enhanced as the listener becomes accustomed to the talker's voice. In particular, we wondered whether hard words that were presented later in a transcription session would be more accurately transcribed than hard words presented earlier in the session. In other words, we were interested in seeing whether listener-talker adaptation might compensate for the processing difficulties introduced by the lexical confusability factor.

Figure 2 shows the percent correct transcription scores for the easy and hard words in the first quartile (Q1) and fourth quartile (Q4) of the transcription sessions at the fast (upper panel), medium (middle panel), and slow (bottom panel) speaking rates. In each case the first and fourth quartiles were

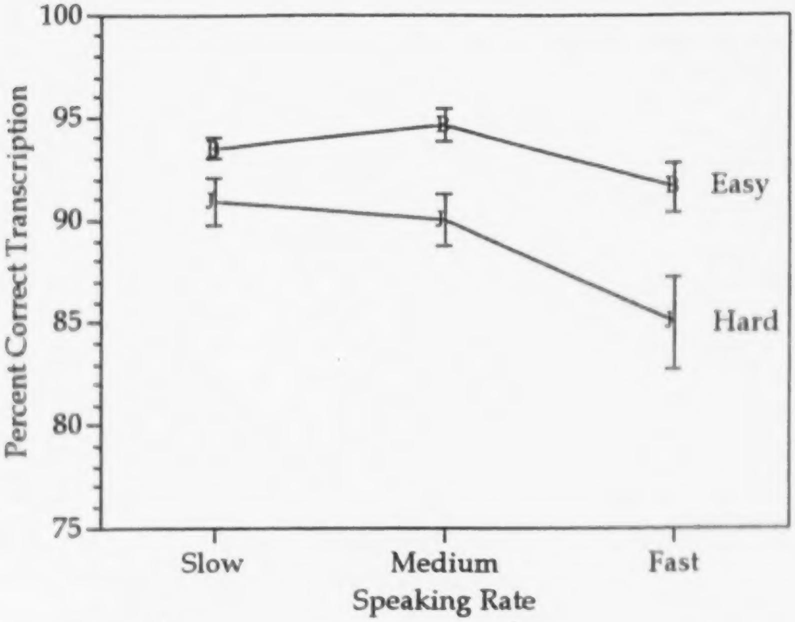


Figure 1. Transcription accuracy for the easy and hard word lists at slow, medium, and fast speaking rates.

taken as the first and last 38 words presented to the listeners, respectively. As shown in Figure 2, hard words presented in the last quartile were generally more accurately transcribed than hard words presented in the first quartile at all three speaking rates. In contrast, there was no noticeable difference between easy words presented in the first and fourth quartiles at all three speaking rates. Separate ANOVA's for each speaking rate showed that for all three rates there was a main effect of quartile, such that the Q4 intelligibility scores were higher than the Q1 intelligibility scores. There was also a main effect of lexical discrimination, such that easy words had higher intelligibility scores than hard words. Furthermore, the quartile by lexical category interaction was significant. Post-hoc tests showed that at all three speaking rates the Q4-Q1 difference was significant for the hard words, but not for the easy words.

Insert Figure 2 about here

These data indicate that as the listener becomes accustomed to the talker's voice and articulatory patterns, the intelligibility difficulty introduced by the lexical characteristics of hard words relative to easy words is "neutralized" to a large extent. Furthermore, a comparison of the first and fourth quartile intelligibility scores across the three speaking rates (see Table 2) showed that the intelligibility of fast rate words in the fourth quartile (mean = 89.67%) approached the intelligibility scores for the slow and medium rate words in the first quartile (means = 90.80% and 90.05%, respectively). In other words, the listener's experience with the talker's speech compensated for the intelligibility difficulty introduced by the fast speaking rate. In general, this pattern of results suggests that listener-talker adaptation is an important factor that interacts with other talker- and item-related factors, such as speaking rate and lexical characteristics, in determining the overall intelligibility of normal speech by normal listeners.

TABLE 2.

Mean intelligibility scores for each speaking rate in the first and fourth quartile.

	First Quartile	Fourth Quartile
Slow	90.80	92.90
Medium	90.05	93.04
Fast	85.98	89.67

Discussion

The primary goal of this study was to examine the combined effects of various talker-, item-, and listener-related factors on normal speech intelligibility by normal listeners. Results showed a strong effect of lexical discrimination (easy words had higher intelligibility scores than hard words), and a strong effect of speaking rate (slow and medium rate words had higher intelligibility scores than fast rate words). Furthermore, we observed a complex relationship between the various factors whereby the difficulties imposed by one factor, such as a fast speaking rate or an inherently difficult lexical item, could be overcome by the advantage gained by the listener's experience with the speech of a particular talker.

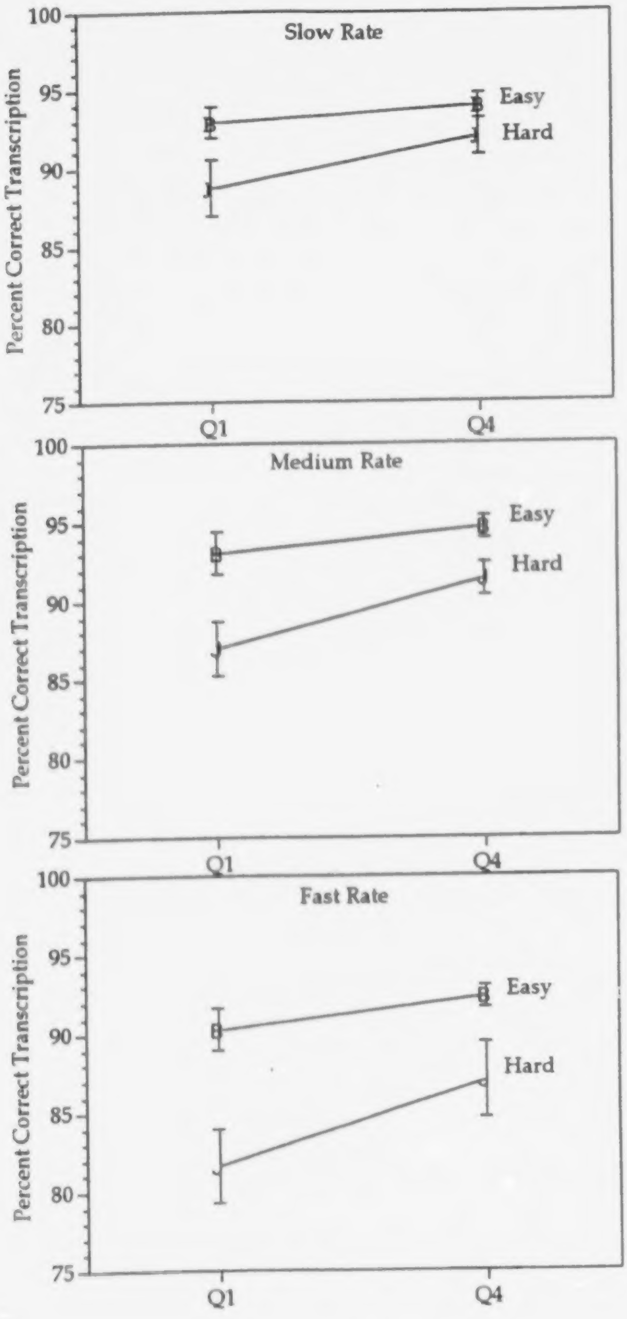


Figure 2. Comparison of transcription accuracies for the easy and hard words presented in the first (Q1) and fourth (Q4) quartile of the transcription sessions at the fast (upper panel), medium (middle panel), and slow (lower panel) speaking rates.

In our investigation of the factors that affect normal speech intelligibility, the present study focused on factors that are related to capabilities unique to speech perception and spoken-language processing. For example, the acoustic-phonetic changes that typically occur as a consequence of a change in speaking rate are directly related to the sound system of the language which imposes limits on the relative expandability and compressibility of various acoustic-phonetic elements. Similarly, the distinction between easy and hard words is directly related to the structure of the lexicon as a whole. Finally, the type of perceptual adaptation that we observed on the part of the listener to the speech patterns of the talker is directly related to the listener's knowledge of the range of possible within-talker variability given the phonetic requirements of the language. Thus, while we do not mean to minimize the importance of basic psychoacoustic capabilities for auditory perception, including speech perception, our focus has been on the higher-level cognitive and linguistic capabilities that are essential for "robust" speech perception and spoken language processing. This focus reflects our concern with the vulnerability of these capabilities in other populations, such as hearing impaired children and adults, non-native listeners, and the elderly.

As we gain a deeper understanding of the operations that are involved in normal speech perception we can begin to develop new tests, and ultimately new training procedures, that focus directly on the complex cognitive and linguistic capabilities that are critical for robust speech perception. Based on the findings of the present study in conjunction with those of other studies reviewed in the introduction, we can delineate several factors that a sensitive robust test of speech perception and spoken language processing should attempt to address. First, the test should examine how listeners cope with stimulus sets that incorporate a high degree of variability due to, for example, multiple talkers and speaking rates. Second, the test should investigate the extent to which listeners perceive words in the context of other words in the lexicon, that is, the extent to which listeners display evidence of having developed a phonetically structured lexicon in long-term memory. Finally, the test should assess the listeners ability to compensate for stimulus-related difficulties by taking advantage of consistent aspects of the speech signal, such as the listener-talker adaptation that we observed in the present data. Such tests are already under development for use with various clinical populations (Kirk et al., 1995; Sommers et al., 1997; Sommers, 1997; Kirk et al., in press). We expect that further development of speech perception assessment instruments that take these factors into account will find a wide range of highly beneficial clinical and research applications.

References

Bradlow, A. R., Torretta, G. M., & Pisoni, D. B. (1996). Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics. *Speech Communication,* 20, 255-272.

Hernandez, L. R. (1995). Current computer facilities in the Speech Research Laboratory. *Research on Spoken Language Processing, Progress Report* , 20, 389-394, Indiana University, Bloomington, IN.

Kirk, K. I., Pisoni, D. B., & Miyamoto, R. C. (In press). Effects of stimulus variability on speech perception in listeners with hearing impairment. *Journal of Speech, Language, and Hearing Research.*

Kirk, K. I., Pisoni, D. B., & Osberger, M. J. (1995). Lexical effects on spoken word recognition by pediatric cochlear implant users. *Ear and Hearing,* 16, 470-481.

Kucera, F. & Francis, W. (1967). *Computational analysis of present day American English.* Providence, RI: Brown University Press.

Luce, P. A. (1986). Neighborhoods of words in the mental lexicon. *Research on Speech Perception, Technical Report No. 6*, Indiana University, Bloomington, IN.

Luce, P. A., Pisoni, D. B., and Goldinger, S. D. (1990). Similarity neighborhoods of spoken words. In G. Altmann (Ed.), *Cognitive models of speech processing: Psycholinguistics and computational perspectives.* Cambridge, MA: MIT Press.

Luce, P. A. & Pisoni, D. B. (In press). Recognizing spoken words: the Neighborhood Activation Model. *Ear and Hearing.*

Mullennix, J. W., Pisoni, D. B., & Martin, C. S. (1989). Some effects of talker variability on spoken word recognition. *Journal of the Acoustical Society of America,* 85, 365-378.

Nusbaum, H. C., Pisoni, D. B., & Davis, C. K. (1984). Sizing up the Hoosier mental lexicon: Measuring the familiarity of 20,000 words. *Research in Speech Perception, Progress Report* , 10, 357-376, Indiana University, Bloomington, IN.

Nygaard, L. C., Sommers, M. C., & Pisoni, D. B. (1994). Speech perception as a talker-contingent process. *Psychological Science,* 5, 42-46.

Pisoni, D. B., Nusbaum, H. C., Luce, P. A. and Slowiaczek, L. M. (1985). Speech perception, word recognition and the structure of the lexicon. *Speech Communication*, 4, 75-95.

Sommers, M. S. (1997). Stimulus variability and spoken word recognition. II. The effects of age and hearing impairment. *Journal of the Acoustical Society of America,* 101, 2278-2288.

Sommers, M. S., Kirk, K. I., & Pisoni, D. B. (1997). Some considerations in evaluating spoken word recognition by normal-hearing, noise-masked normal-hearing, and cochlear implant listeners. I: The effects of response format. *Ear and Hearing,* 18, 89-99.

Sommers, M. S., Nygaard, L. C. & Pisoni, D. B. (1994). Stimulus variability and spoken word recognition: I. Effects of variability in speaking rate and overall amplitude. *Journal of the Acoustical Society of America*, 96, 1314-1324.

Torretta, G. M. (1995). The easy-hard word multi-talker speech database: An initial report. *Research on Spoken Language Processing, Progress Report* , 20, 321-334, Indiana University, Bloomington, IN.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 21 (1996-1997)
Indiana University

Audio-Visual Speech Perception Without Traditional Speech Cues: A Second Report[1]

**Robert E. Remez,[2] Jennifer M. Fellowes,[2] David B. Pisoni,
Winston D. Goh, and Philip E. Rubin[3]**

*Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405*

---

[1] This research was funded by NIH NIDCD Research Grant DC00111 to Indiana University and Research Grant DC00308 to Barnard College.

[2] Department of Psychology, Barnard College, New York, NY.

[3] Haskins Laboratories, New Haven, CT.

R.E. REMEZ, J.M. FELLOWES, D.B. PISONI, W.D. GOH, AND P.E. RUBIN

Audio-Visual Speech Perception
Without Traditional Speech Cues: A Second Report

Abstract. Theoretical and practical motives alike have prompted investigations of multimodal speech perception. Theoretically, such studies lead the explanation of perceptual organization beyond the familiar modality-bound accounts deriving from Gestalt psychology. Practically, existing perceptual accounts fail to explain the proficiency of multimodal speech perception using an electrocochlear prosthesis for hearing. Accordingly, our research sought improved measures of audiovisual integration of videotaped faces and selected acoustic constituents of speech signals with an acoustic signal that departs from the natural spectral properties of speech. A single sinewave tone accompanied a video image of an articulating face; the frequency and amplitude of the phonatory cycle or of one of the lower three oral formants supplied the pattern for a sinewave signal. Our results showed a distinct advantage for the condition pairing the video with a sinewave replicating the second formant, despite its unnatural timbre and its presentation in acoustic isolation from the balance of the speech signal.

Introduction

How does the perceiver find the speech signal amid an uninterrupted flux of sensory activity? The customary answer to this question discusses the principles of perceptual organization intrinsic to each of the sensory modalities, following Wertheimer (1923). In essence, two classes of general principle, visual and auditory, are available to apply to speech, and few clear proposals aim to explain perceptual organization when the listener also looks at the talker. There is hardly any doubt that multimodal perceptual organization does actually occur, and a small but sturdy literature describes perceptual phenomena that falsify the description of post-perceptual integration (for instance, Green & Miller, 1985).

Research on sinewave replicas of speech has been singular in promoting a different approach to the problem of organization than is customary (Julesz & Hirsh, 1972). Although a tonal analog of a speech signal is intelligible (Remez, Rubin, Pisoni & Carrell, 1981), it lacks the typical acoustic manifestations of vocal sound production. This fact arguably demonstrates the limitations of general auditory accounts of perceptual organization (Bregman, 1990) and probabilistic accounts of speech perception that rest on likely correspondence of signal element and phonetic segment (Massaro, 1994). Instead, findings with sinewave replicas of speech provide evidence of an alternative account of perceptual organization based on susceptibility to the unique spectrotemporal characteristics of a phonologically modulated source of sound (Remez, Rubin, Berns, Pardo & Lang, 1994). Perceptual organization, in this view, exploits a perceiver's sensitivity to patterned spectra, in contrast to the piecemeal assessment of elemental details of the acoustic stream warranted by prior accounts. These findings belong to an emerging class of reports about speech which note the integration of sensory elements despite detailed dissimilarity in their physical properties.

Multimodal Perceptual Organization of Speech

Although the perceptual organization of speech in an auditory system appears well characterized from a consideration of sinewave replicas, it also seems that multimodal speech perception exhibits some of the main characteristics identified by this line of research. Principally, the organization of visual and auditory inflow in a bimodal case of speech perception appears to conjoin stimulation from the modalities preliminary to an analysis of the patterned unimodal sensations. A clear case of this phenomenon is seen in a report by Green and Miller (1985) who observed that the identification of syllables in an auditory voicing series was a function of silent visual information about the rate of articulation. Had the rate information been specified acoustically, the outcome of the tests would have been explained agreeably as evidence of a kind of context for analysis of the spectrotemporal acoustic pattern that varied to evoke an impression of voiced and voiceless consonants. In the bimodal case, though, there is no perceptual function readily available to explain the lability of phonetic analysis to a combination of visual and auditory stimulation. Although Welch and Warren (1980) held that multimodal integration might depend on a common spatial locus for sound and sight, this premise falsely predicts failure of dichotic fusion of speech (Broadbent & Ladefoged, 1957; Remez et al., 1994). The finding of Green and Miller (1985) is especially provocative considering that their subjects perceived a phonetic contrast that depends on fine resolution of sequential patterning, indicating that sensory streams are combined in a manner that is temporally veridical.

Two studies set the question of multimodal organization directly. In one, by Breeuwer and Plomp (1985), speechreading was supplemented with pure tones modulated at the frequencies of the first and the second formant. Subjects transcribed the audiovisual conditions relatively poorly, as if the tone analogs of the formants were barely fused with the visual impression of the articulating face. In contrast, Bernstein et al. (1992) used an acoustic or tactile presentation of the frequency band of the first or the second formant, and observed great benefit to speechreading of either F1 or F2 in a concurrent auditory signal, and an enhancement of speechreading with a tactile vocoder driven by the variation in the frequency region of F2. Clearly, a tone reproducing the frequency variation of the second formant cannot both be effective and ineffective in audiovisual presentation.

The Problem of the Second Formant

One clue about the cause of the different effects is the method used in each study to analyze the formant pattern. Breeuwer and Plomp argued that accurate assessment of formant frequency cannot be accomplished in real time. Their goal of assessing the prospects of an instrumental aid to perception required them to use existing signal processing technology, and they adopted linear prediction with minimal correction to determine formant values for voiced speech only. Although we can be confident that the temporal alignment of the resulting frequency modulated tones was accurate, the unvoiced formant values were simply missing, and other samples were unquestionably erroneous due to interpolation when the LPC analysis simply failed. This was not a completely satisfactory test of the perceptual organization of time-varying auditory and visual stimulation during speechreading because the auditory values were probably misleading.

The group led by Bernstein used the labels F1 and F2 to describe the patterns produced by their vocoders, but in actuality they used the output of stationary filter banks that approximated the range over which the first or second formant frequency excursions occurred. For F1, this was 75-900 Hz; for F2, it was 975-2625 Hz. It is likely, therefore, that the nominal F2 often included the third formant, and it is possible that the nominal F1 contained the second formant for some back vowels and labial consonants. This method fell short of an exact test of the perceiver's disposition to organize visual displays of the face and individual formant bands in speech perception.

Our own recent attempt to provide a clear resolution to this multimodal problem of integrating the second formant and the visual impression of a talker was less than successful (Saldaña, Pisoni, Fellowes & Remez, 1996). We used single tones from sinewave utterance replicas in combination with a video display of the face, and found that the greatest benefit to normal hearing subjects occurred when the moving image

BEST COPY AVAILABLE

of the face was combined with the tone analog of the second formant. Other multimodal conditions included tone analogs of the first formant, of the F0 pattern, and a noise band modulated in amplitude according to the overall energy in the signal. The finding of greatest benefit attending the audiovisual combination of the second formant analog and the face occurred without natural timbre, of course. This result is consistent with prior findings by Bernstein et al. (1992), and suggests that accurate estimates of the frequency of the second formant produce benefits in the multimodal case, contrary to Breeuwer and Plomp (1992) who used uncorrected linear-prediction estimates. However, the performance levels in our earlier study were low (Saldaña et al., 1996). In a control condition using complete tonal replicas based on the utterances of this talker, average performance did not exceed 35% of syllables correct, whereas more typical performance on sinewave sentences approaches 80% correct. The cause, we suspect, was the talker, whose speech was unpredictably difficult for our listeners, a possibility which we verify in the present study.

The Present Test of Multimodal Integration

To conduct a fairer test of multimodal coherence, we based our audio-visual presentation on the speech of a demonstrably intelligible talker (Bradlow, Torretta & Pisoni, 1996) to attempt to bring test performance off the floor, thereby resolving any differential effects of the single tones in combination with the video presentation. On the basis of the performance in this dataset, we recruited an individual to read a sentence list while video and audio signals are sampled. The natural speech was converted to sinewave replicas, and multimodal coherence was assessed in transcription tests combining the visual presentation with the tonal analog of the first, second or third formant; and with a tone replicating the pattern of the fundamental frequency of phonation.

Test Materials.

An adult female whose natural speech was verified as highly intelligible produced utterances that were sampled for video and audio reproduction. Ten sentences were selected from the dataset of Bradlow et al. (1996) and were spoken from a list.

Insert Figure 1 about here.

Formant center frequency and amplitude were estimated interactively by comparing discrete Fourier spectra and linear prediction estimates; frequency of phonation was estimated from a narrow-band Fourier representation of the spectrum. Frequency and amplitude values for F0, F1, F2, and F3 were converted to four single-tone time-varying sinusoids using a software synthesizer (Rubin, 1980). The computed sinusoidal waveforms were combined and synchronized with the video, and presented on-line to listeners in individual testing carrels.

Procedure.

Thirty-eight test takers were assigned randomly to one of four audiovisual conditions, Video+ToneF0, Video+ToneF1, Video+ToneF2, and Video+ToneF3. Each session began with a sequence of eight four-tone sinewave sentences, presented to give the subjects a brief chance to adjust to the odd timbre of sinewave signals. These test items were based on the speech of one of the authors, and did not duplicate any of the test sentences used in the multimodal conditions.

Following the familiarization sequence, an audiovisual condition began. Each sentence was repeated five times, after which the subject was cued to write a faithful rendition of the message in a

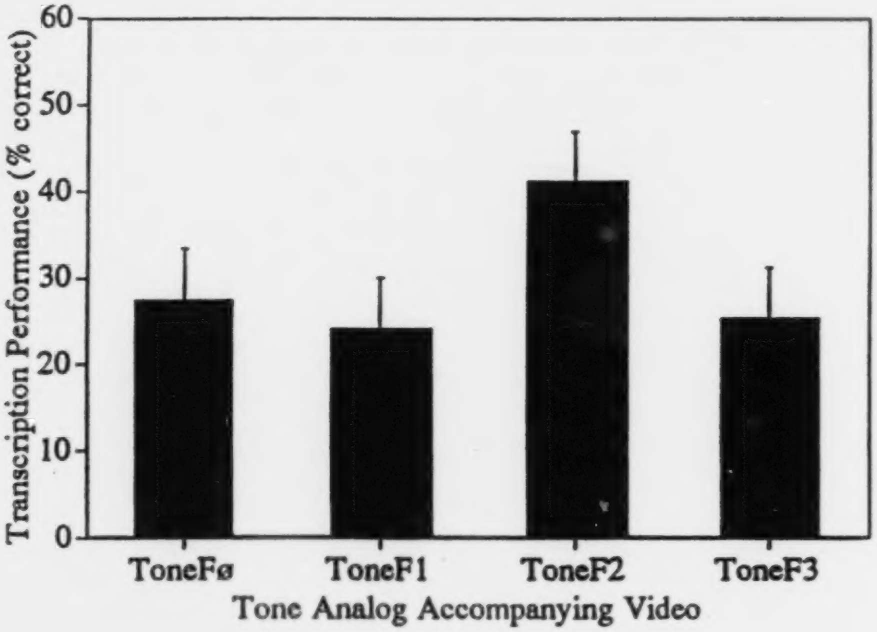


Figure 1. Results of a test of audiovisual speech perception with tone analogs of speech. Each bar shows the group performance with a different signal component. Error bars represent the confidence region for a post hoc means test (Tukey, $\alpha = .05$).

specially prepared test booklet. A warning tone also occurred before the start of a new sentence block, to alert the subject to finish writing and to look at the video monitor.

At the conclusion of the audiovisual test conditions, a repetition of the initial set of eight audio sinewave sentences occurred. This served as a check on the absolute ability of subjects to derive phonetic impressions from sinewave signals. This sort of assessment has been necessary due to the immunity to the phonetic properties of sinewave signals of a substantial subset of volunteer subjects, though none of the thirty-eight participants in this sample was excluded on such grounds (see Remez et al., 1994).

Results.

A transcription provided by a participant in an audiovisual test was scored by tallying the percent of the syllables in each sentence that had been transcribed correctly, following established procedure (Remez et al., 1981). Each subject contributed ten values, one for each of the test sentences, to a one-way analysis of variance of the effect on transcription performance of the four single tones combined with the video sample. Despite the small size of the groups (10 subjects in the first and second groups, 9 subjects in the third and fourth groups), the effect of the tone manipulation was found to affect performance [$F(3, 34) = 5.99$, $p < .002$]. The group performance in the four test conditions is shown in Figure 1. It is plain to see that the tone analog of the second formant, in combination with the video samples, produced performance that was significantly better than that which we observed in the three other tones.

Discussion

The pattern of results, in which the tone analog of the second formant combined more effectively with the video samples than the other single tones that we tested, suggests an interpretation of the three studies that had set the specific empirical problem for us. First, although we derived test materials from visual and acoustic samples of the speech of an intelligible talker, the performance levels here replicated the pattern of our earlier observation (Saldaña et al., 1996). A tone exhibiting the pattern of F2 made a more effective acoustic accompaniment to the video samples than did the tone analogs of the other formants or the fundamental, and our findings show that there is no second best; performance in three conditions with the other tones was equal.

On the reports of prior research, we might have expected the analog of the first formant (Bernstein et al., 1992) or of the fundamental frequency of phonation (Rosen, Fourcin & Moore, 1981) to combine readily with the video samples in evoking an impression of the linguistic message. Differences in the linguistic test materials are important to consider, because the unforgiving sentences that we used here may have inadvertently suppressed the differences in effectiveness of tones other than ToneF2. Nonetheless, for multimodal perceptual organization in which the auditory component lacks the timbre of natural speech, it is safe to conclude that the unique effectiveness of the analog of the second formant is established more solidly by these results.

Second, a comparison is also appropriate of this multimodal circumstance to the effects of dichotic presentation of sinusoidal sentence components (Remez et al., 1994). In that study, one ear received an isolated tone analog of the second formant, the other ear received the balance of the tones composing the sentence replica. Transcription performance for the concurrent presentation well exceeded the performance predicted by assaying the intelligibility of the components separately. The same kind of concurrent benefit is likely to have obtained in the multimodal case presented here. Neither the video samples of the talker's face nor the impressions evoked by the analog of the second formant are known to elicit accurate or definite impressions of the phonetic properties of a message. Yet, in analogy to the dichotic case, the concurrent presentation allowed listeners to organize the multimodal inflow and to transcribe about half of the syllables correctly in a difficult set of sentences.

Coincidentally, the performance levels are roughly the same for dichotic sinewave sentences and multimodal sinewave sentences. A clue to perceptual organization may reside in this similarity. If a synthetic second formant exhibiting natural timbre is more effective multimodally than a tone analog of F2, this would indicate that organizational functions may be contingent on short-term spectrum in some instances. Alternatively, prolonged exposure to sinewave signals may acclimate subjects to the anomalous timbre of the sinewave voice, and such a procedure may be seen as improvements in performance due solely to perceptual tuning.

Last, the principle that we proposed to explain the dichotic combination of acoustic information was based on susceptibility to the spectrotemporal patterns of an acoustic signal independent of its superficial properties. Specifically, in the case of speech the principle is evidently matched, albeit abstractly, to the physical structure of vocal resonators and the functional organization of phonologically governed articulation. Because a sinewave differed physically from the acoustic signal elements it replicated in coarse grain, no perceptual evaluation of elementary "speech cues" alone would accommodate the finding.

To accommodate the multimodal case of speech, the organizational principle satisfied by the auditory and the visual inflow must be still more abstract. By such means the perceiver treats the sensory inflow as information about a unitary event distributed across multiple modalities: auditory, visual, vibrotactile, haptic orosensory, and motoric. Our search for a description of this system of linguistic contrasts and multiple sensory projections may eventually explain why the frequency excursions of the second formant combine so readily with the visual presentation of the articulating face.

References

Bernstein, L.E., Coulter, D.C., O'Connell, M.P., Eberhardt, S.P., & Demorest, M.E. (1992). Vibrotactile and haptic speech codes. Lecture presented at the *Second International Conference on Tactile Aids, Hearing Aids, & Cochlear Implants*. Royal Institute of Technology, Stockholm, Sweden, June 9-11, 1992.

Bradlow, A.B., Torretta, G.M., & Pisoni, D.B. (1996). Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics. *Speech Communication*, 20, 255-272.

Breeuwer, M., & Plomp, R. (1985). Speechreading supplemented with formant-frequency information from voiced speech. *Journal of the Acoustical Society of America* 77, 314-317.

Bregman, A.S. (1990). *Auditory Scene Analysis*. Cambridge: MIT Press.

Broadbent, D.E., & Ladefoged, P. (1957). On the fusion of sounds reaching divergent sense organs. *Journal of the Acoustical Society of America*, 29, 708-710.

Green, K.P., & Miller, J.L. (1985). On the role of visual rate information in phonetic perception. *Perception & Psycnophysics*, 38, 269-276.

Julesz, B., & Hirsh, I.J. (1972). Visual and auditory perception: An essay of comparison. In E.E. Denes & P.B. Denes (Eds.), *Human Communication: A Unified View* (pp. 283-340). NY: McGraw-Hill.

Massaro, D.W. (1994). Psychological aspects of speech perception: Implications for research and theory. In M.A. Gernsbacher (Ed.), *Handbook of Psycholinguistics* (pp. 219-263). NY: Academic Press.

Remez, R.E., Rubin, P.E., Berns, S.M., Pardo, J.S., & Lang, J.M. (1994). On the perceptual organization of speech. *Psychological Review*, 101, 129-156.

Remez, R.E., Rubin, P.E., Pisoni, D.B., & Carrell T.D. (1981). Speech perception without traditional speech cues. *Science*, 212, 947-950.

Rosen, S.M., Fourcin, A.J., & Moore, B.C.J. (1981). Voice pitch as an aid to lip-reading. *Nature*, 291, 150-152.

Rubin, P.E. (1980). Sinewave synthesis. Internal memorandum, Haskins Laboratories, New Haven, CT.

Saldaña, H.M., Fellowes J.M., Remez, R.E., & Pisoni, D.B. (1996) Audio-visual speech perception without speech cues: A first report. In D.G. Stork and M.E. Hennecke (Eds.), *Speechreading by Man and Machines: Models, Systems and Applications* (pp. 145-151). Berlin: Springer-Verlag.

Welch, R.B., & Warren, D.H. (1980). Immediate perceptual response to intersensory discrepancy. *Psychological Bulletin*, 88, 638-667.

Wertheimer, M. (1923). Untersuchungen zur Lehre von der Gestalt, II. *Psychologische Forschung*, 4, 301-350. [Reprinted in translation as "Laws of organization in perceptual forms," in W.D. Ellis (Ed.), *A Sourcebook of Gestalt Psychology* (pp. 71-88). London: Routledge & Kegan Paul, 1938.]

III. Instrumentation and Software

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 21 (1996-1997)
Indiana University

The Hoosier Audiovisual Multi-Talker Database[1]

Sonya Sheffert, Lorin Lachs and Luis R. Hernández

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405

---

[1] This research was supported by NIH-NIDCD Research Grant DC00111 and NIH-NIDCD Training Grant DC00012 to Indiana University Bloomington.

The Hoosier Audiovisual Multi-Talker Database

Abstract. This report describes the Hoosier Audiovisual Multi-Talker Database, a 3000 word multimodal video database developed at the Indiana University Speech Research Laboratory. The corpus consists of ten adult talkers (five male and five female) producing 300 familiar monosyllabic English words. Each spoken word is presented as a dynamic full-motion color movie. The database also includes information about the lexical characteristics and the intelligibility of each word.

Objectives of the Database

The objectives of our research program on multimodal perception are to: 1) investigate the perception and integration of auditory and visual information during multimodal language processing, 2) explore the nature of the memory traces for spoken words produced by different speakers, 3) assess the effects of optical and auditory talker information on speaker normalization, 4) examine the relationship between implicit and explicit memory for voices, faces, and spoken words, 5) investigate the effects of talker-specific characteristics on multimodal speech intelligibility, and 6) determine whether structural characteristics of the mental lexicon contribute to listener's ability to use multimodal information during the course of word recognition. In order to achieve these goals, it was necessary to create a large digital database of spoken words.

Organization of the Database

The organization of the database is based on lexical neighborhoods, or collections of words characterized by acoustic-phonetic similarity (Luce, 1986; Luce & Pisoni, 1998). There are 150 words designated as "easy", and 150 as "hard". Easy words reside in sparse neighborhoods populated by only low frequency words, whereas hard items reside in dense neighborhoods shared by many high frequency words.

The database is organized around acoustic-phonetic similarity because previous research has demonstrated that these two diminsions, frequency and similarity play an important role in spoken word identification. Target items from low-density, low-frequency neighborhoods are identified faster and more accurately than hard words, presumably because there is less lexical competition among the items. Differences in the identifiability of easy and hard words have been demonstrated in normal listeners (Luce, 1986; Luce & Pisoni, 1998), in hearing impaired listeners and patients with cochlear implants (Sommers, Iler-Kirk, Pisoni & Osberger, 1993), young children (Charles-Luce & Luce, 1990) and older adults (Sommers, 1996). The development of Hoosier Audiovisual Multi-Talker Database allows us to extend the findings by addressing the effects of lexical confusability on multimodal language processing.

Because each word was produced by ten different speakers (five male and five female), the database can also be organized by talker. The use of several talkers permits us to build on a growing literature demonstrating that speech perception and spoken word recognition are influenced by stimulus variability and talker familiarity (for a review, see Pisoni, 1993).

The tokens in the databases are dynamic full-motion color movies rather than as static photographs. This is important for exploring issues concerned with the integration of multimodal information; that is, rather than having audio information which is tied to static visual information arbitrarily, the database contains two tracks of information from different sensory modalities which are lawfully tied to one another. This will allow users of the database to investigate questions concerning the multimodal encoding and processing of speech in a more ecologically valid manner.

Creating the Hoosier Audiovisual Multi-Talker Database

Methods

Subjects

Ten different talkers (five males and five females) were recruited from the Indiana University community. The age of the talkers ranged from 18 years to 32 years, and they represented a range of geographical areas in the Midwest: Chicago (N=4), Indiana (N=1), Iowa (N=1), Oklahoma (N=2), St. Louis (N=1) and Wisconsin (N=1). None of the talkers wore glasses. One talker had a small mustache and beard at the time of the recordings. All talkers were Caucasian.

Stimulus Materials

Each talker was videotaped while producing 300 monosyllabic CVC words. The words were selected from the Hoosier Mental Lexicon (Nusbaum, Pisoni, & Davis, 1984), which is based on a 20,000 word on-line dictionary. The HML database provides information on word frequency (Kucera & Francis, 1967), word familiarity (Nusbaum et al., 1984), neighborhood density and neighborhood frequency. Together, these variables allowed one to specify the relative degree of confusability among the items, and to separate the words into "easy" and "hard" items.

Only highly familiar words (6 or greater on a 7-point scale) were selected for the Hoosier Audiovisual Multi-Talker Database. Neighborhood density or the number of phonetically similar words or neighbors of a target item, was determined by a one-phoneme substitution, addition, and deletion metric (Greenberg & Jenkins, 1964). Mean neighborhood density was 18 for the easy items and 24 for the hard items. Neighborhood frequency, or the average frequency of all the items within a target neighborhood, was obtained from Kucera and Francis (1967). Mean neighborhood frequency was 41 per million for the easy items and 251 per million for the hard items. A summary of the lexical characteristics of the items is displayed in Table 1.

Table 1.

Lexical characteristics of the easy and hard words.

	Easy	Hard
Mean Lexical Frequency	319	28
Mean Familiarity	7	7
Mean Lexical Density	18	24
Mean Neighborhood Frequency	41	251

Talkers were videotaped in a sound-attenuated professional recording studio using an 8mm professional SVHS Canon video camera. Each word was presented in isolation on a CRT screen at a fixed citation rate of 1 word every 3 seconds. Each talker received a different random ordering of the words.

During the videotaping, all the speakers wore a solid black shirt. Talkers were instructed to speak clearly and naturally at a normal conversational rate. They were told to look directly into the camera while assuming a neutral facial expression and to avoid any extraneous head to body movement. They were also instructed to begin and end each utterance with their mouth closed. Words which were mispronounced or accompanied by extraneous movement were re-recorded.

The video images were captured, digitized and segmented using a commercial software package (Adobe Premiere) installed on a Macintosh Quadra 950. The video equipment was designed to record and playback full-motion video at 30 frames per second (NTSC standards). The audio signal was digitally sampled at 22kHz with 16-bit resolution. The video was digitized at 30 fps, with 24-bit resolution at 640 by 480 pixel size. Each word was made into a free standing Quicktime movie using Adobe Premier's movie making function. The movies were created in Radius format and converted to TARGA format in order to maintain compatibility with the latest video hardware. The audio signal for all tokens across talkers was equated for root mean square amplitude (RMS). All movies were leveled at 54 dB. The resulting leveled movies were stored on digital optical disks.

Each token is approximately 2 seconds long. Every spoken word is buttressed by approximately .5 seconds of silence during which the talker's mouth was closed. The overall integrity of each movie was assessed by two trained listeners (a phonetician and a psychologist). Each token was screened for the presence on any of the following errors: Acoustic distortion, background noise, ambiguous or incorrect pronunciation, nonspeech mouth sounds; unusual lip movement, unusual eye movement, and visual distortion. The final version of the movies are presented as a full-screen image (640 x 480) in 24-bit color and presented on a high resolution 17'' monitor.

Speech Intelligibility

Future users of the database may need data on the intelligibility of each of the tokens. Therefore, each stimulus was tested for ease of perceptual identification. Results showed that the majority of tokens in the database were highly intelligible with little variability between talkers. We report only the AV intelligibility in this report. Assessment of the intelligibility of these stimuli in the visual- and audio-only contexts is currently in progress.

Subjects

100 Indiana University undergraduates participated in return for partial course credit in an introductory psychology course or for five dollars. All subjects had normal hearing, were native English speakers, and reported no history of speech or hearing disorders at the time of testing.

Materials

One Macintosh Quadra 950, one Macintosh PowerPC 7100, and three PowerComputing 180s, each with a 17" Sony Trinitron Monitor, were used to present the video displays of the stimuli. The stimuli consisted of the movies from all ten talkers.

Procedure

A computer program was used to control stimulus presentation and collect subject response. Each subject was presented with a randomly ordered set of movies; each set of movies consisted of all the tokens spoken by a particular talker. Stimuli were presented using 17" monitors and BeyerDynamics DT-100 stereophonic headphones at a loudness of 75 dB/SPL. Before the presentation of the first stimulus, subjects were given a set of typed instructions explaining the task and procedures. Listeners were informed that they would see a series of movies in which a person would speak a single English word. Subjects were informed that they would be hearing each word only once. After each stimulus, subjects were required to identify the word by typing its spelling on the keyboard. Subjects were instructed that the next movie would not be presented until they pressed the RETURN key. They were also reminded to take time to make sure that the response they typed was the response which they intended to make before entering it. Each response was then collected in a text file which contained the name of the movie, it's order in the presentation, and the subject's response. Each subject, therefore, had his/her own logfile.

Data Analysis

All logfiles from subjects who viewed a particular talker were analyzed in tandem. The intelligibility data program scanned the responses made across subjects for each movie. Responses that matched the intended response were scored as correct. Any strings which were homophonic with the intended response or any strings which matched the correct response except for obvious typos were also accepted as correct. All other responses were marked as incorrect identifications of the target word. The data were analyzed for the effects of three factors on intelligibility: talker, easy/hard lexical classification, and word.

Results

The overall intelligibility of the words, expressed as the percent correct identification of all words from all talkers, was 98.57%. The mean intelligibility of each word across talkers was also calculated and a table of these scores was constructed for further reference. This table will allow users of the database to have readily available statistics on the intelligibility of the words contained therein and will allow the use of relative intelligibility as a possible experimental variable in the future. Closer examination of the table reveals several relevant characteristics of the database as a whole. The lowest average intelligibility was 74.09% (for the word "cot"). However, most items produced intelligibility scores that were near ceiling. In fact, examination of the distribution of scores reveals that 99.97% of the words were identified across talkers with greater than 90% accuracy.

Figure 1 shows the intelligibility scores for each talker separated by lexical category. A 2x10 ANOVA (Lexical Category, Talker) performed on the items revealed a main effect of Lexical Category [F(1.3010) = 5.4; p=0.02]. On average, Easy Words were identified 0.5% better than Hard words. However, there was no main effect of talker [F(9,3010) = 0.3051; N.S.], and no significant interaction between these two variables [F(9,3010) = 0.525; N.S.]. Post-hoc Fisher's PLSD pairwise analysis of the talker combinations showed no significant difference between any two specific talkers. In addition, a one-way ANOVA by Sex of the Talker showed no significant effect of this variable [F(1,9) = 0.112; N.S.]. These results are not surprising in light of the fact, noted above, that most of the intelligibility scores were near ceiling levels.

Insert Figure 1 About Here

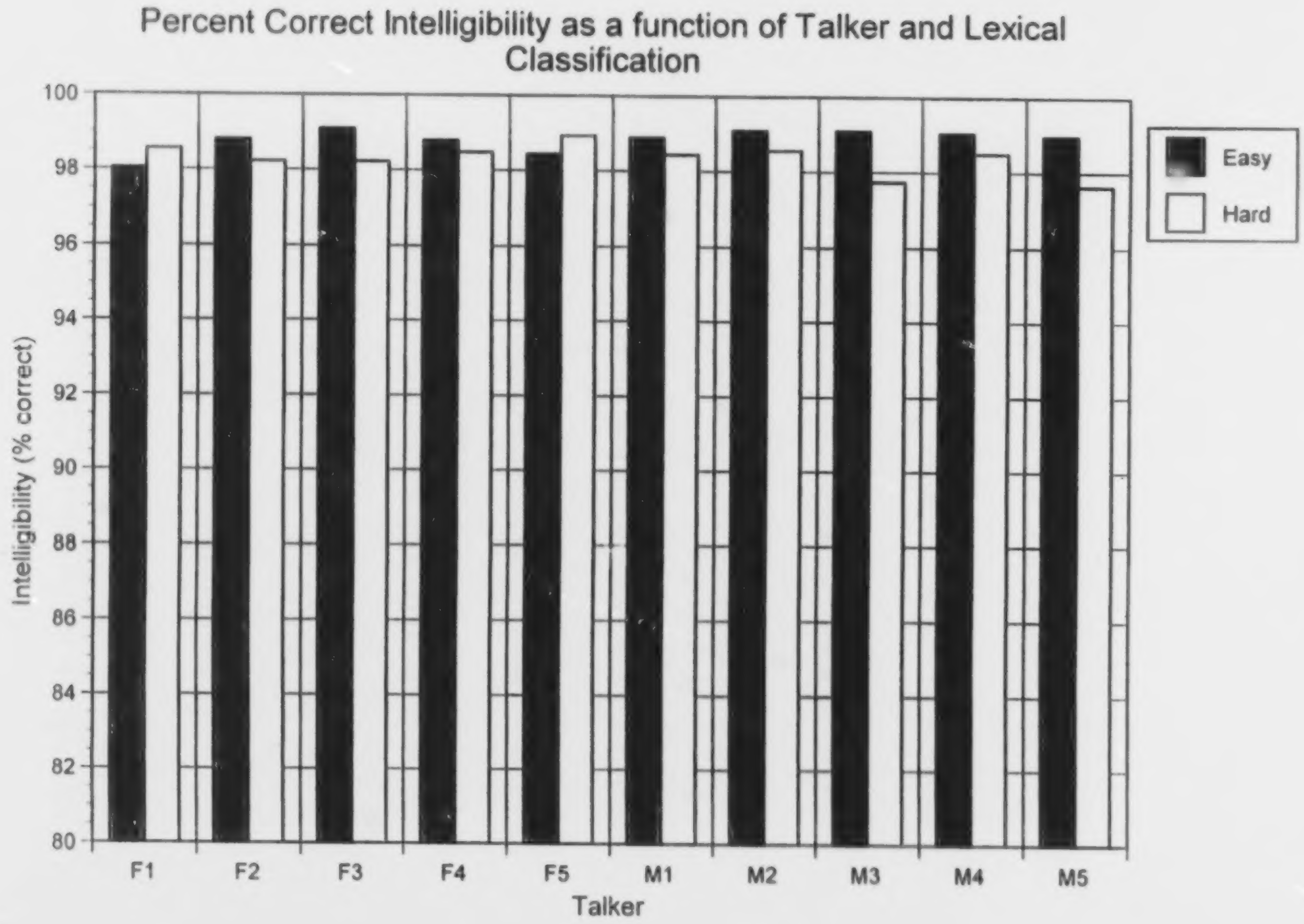


Figure 1. Mean percent correct intelligibility of "Easy" and "Hard" words, displayed as a function of speaker. F1 through F5 refer to the female speakers; M1 through M5 refer to the male speakers.

Conclusion

The development of the Hoosier Audiovisual Multi-Talker Database was completed in two phases. In the first phase, the stimuli were filmed, digitized, leveled and stored. During the second phase, the stimuli were subjected to a test of intelligibility. All in all, we was found that there were no significant differences between stimuli across the talker who spoke them. In addition, there were no significant differences in intelligibility as a function of the words contained in the stimuli. There was, however, a significant difference between stimuli classified as Easy or Hard lexically. Although it may at first glance appear detrimental that differences between stimuli were not found to be significant, it should be remembered that the vast majority of the stimuli contained within the database had intelligibility scores near ceiling. Overall, more than 99% of the tokens were identified with greater than 90% accuracy. This is most likely the reason for a lack of significant effects, but it at the same time attests to the fact that the construction of the Hoosier Audiovisual Multi-Talker Database has produced a large corpus of stimuli which are highly intelligible and are readily available for use in future investigations using audiovisual stimuli.

References

Charles-Luce, J. & Luce, P. A. (1990). Similarity neighborhoods of words in young children's lexicons. *Journal of Child Language,* **17,** 205-215.

Greenberg, J. H. & Jenkins, J. J. (1964). Studies in the psychological correlates of the sound system of American English. *Word*, **20**, 157-177.

Kucera, F. & Francis, W. (1967). *Computational analysis of present-day American English.* Providence, RI: Brown University Press.

Luce, P. A. (1986). Neighborhoods of words in the mental lexicon. *Dissertation Abstracts International,* **47** (12-B, Pt. 1), 5078.

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear & Hearing*, **19**, 1 - 36.

Nusbaum, H. C., Pisoni, D. B. & Davis, C. K. (1984). Sizing up the Hoosier mental lexicon: Measuring the familiarity of 20,000 words. In *Research on Speech Perception Progress Report No.* **10** (pp. 357-377). Bloomington, IN: Speech Research Laboratory, Indiana University.

Pisoni, D. B. (1993). Long-term memory in speech perception: Some new findings on talker variability, speaking rate and perceptual learning. *Speech Communication,* **13,** 109-125.

Sommers, M. S. (1996). The structural organization of the mental lexicon and its contribution to age-related declines in spoken word recognition. *Psychology and Aging,* **11**, 333-341.

Sommers, M. S., Kirk, K. I., Pisoni, D. P & Osberger, M. J. (1993). Some new directions in evaluating the speech perception performance of cochlear implant patients. A first report. In *Research on Spoken Language Processing Progress Report No.* **19** (pp. 271-281). Bloomington, IN: Speech Research Laboratory, Indiana University.

IV. Publications

632

IV. Publications

Articles Published

Bradlow, A.R. (1996). A perceptual comparison of the /i/–/e/ and /u/–/o/ contrasts in English and in Spanish: Universal and language-specific aspects. *Phonetica, 53,* 55-85.

Bradlow, A.R., Pisoni, D. B., Yamada, R.A., & Tohkura, Y. (1997). Training Japanese listeners to identify English /r/ and /l/: IV. Some effects of perceptual learning on speech production. *Journal of the Acoustical Society of America, 101,* 2299-2310.

Bradlow, A.R., Torretta, G.M., & Pisoni, D.B. (1996). Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics. *Speech Communication, 20,* 255-272.

Dinnsen, D.A., Barlow, J.A., & Morrisette, M.L. (1997). Long-distance place assimilation with an interacting error pattern in phonological acquisition. *Clinical Linguistics and Phonetics, 11,* 319-338.

Frisch, S. (1997). Against underspecification in speech errors. In J. Cole (ed.), *Proceedings of the 2nd Mid-Continental Workshop on Phonology. Studies in the Linguistic Sciences 27*(1), 79-97.

Frisch, S. (1997). Review of "Speech: A special code" by A.M. Liberman. *Linguist List* 8.954.

Gierut, J.A., & Morrisette, M.L. (1996). Triggering a principle of phonemic acquisition. *Clinical Linguistics and Phonetics, 10,* 15-30.

Gierut, J.A., Morrisette, M.L., Hughes, M.T., & Rowland, S. (1996). Phonological treatment efficacy and developmental norms. *Language, Speech and Hearing Services in Schools, 27,* 215-230.

Kirk, K.I., Pisoni, D.B. & Miyamoto, R.C. (1997). Effects of stimulus variability on speech perception in hearing impaired listeners. *Journal of Speech, Language, and Hearing Research, 40,* 1395-1405.

Lively, S.E., & Pisoni, D.B. (1997). On prototypes and phonetic categories: A critical assessment of the perceptual magnet effect in speech perception. *Journal of Experimental Psychology: Human Perception and Performance, 23,* 1665-1679.

Meyer, T.A., & Bilger, R. (1997). Effect of set size and method on speech-reception thresholds in noise. *Ear and Hearing, 18,* 202-209.

Pisoni, D.B. (1996). Word identification in noise. *Language and Cognitive Processes, 11,* 681-687.

Ryalls, B.O., & Pisoni, D.B. (1997). The effect of talker variability on word recognition in preschool children. *Developmental Psychology, 33,* 441-451.

Sommers, M.S., & Kewley-Port, D. (1996). Modeling formant frequency discrimination of female vowels. *Journal of the Acoustical Society of America, 99,* 3770-3781.

Sommers, M.S., Kirk, K.I., & Pisoni, D.B. (1997). Some considerations in evaluating spoken word recognition by normal-hearing and cochlear implant listeners I: The effects of response format. *Ear & Hearing, 18,* 89-99.

BOOK PUBLISHED

Chin, S.B. & Pisoni, D.B. (1997). *Alcohol and speech.* San Diego, CA: Academic Press.

BOOK CHAPTERS PUBLISHED

Chin, S.B. (1996). The role of the sonority hierarchy in delayed phonological systems. In T.W. Powell (Ed.), *Pathologies of speech and language: Contributions of clinical phonetics and linguistics* (pp. 109-117). New Orleans, LA: International Clinical Phonetics and Linguistics Association.

Kirk, K.I., Diefendorf, A.O., Pisoni, D.B. & Robbins, A.M. (1997). Assessing speech perception in children. In L.L. Mendel and J.L. Danhauer (Eds.), *Speech perception assessment* (pp. 101-132). San Diego, CA: Singular Press.

Miyamoto, R.T., Kirk, K.I., Robbins, A.M., Todd, S.L., Riley, A.I., & Pisoni, D.B. (1997). Speech perception and speech intelligibility of children with multichannel cochlear implants. In L. Honjo and H. Takahashi (Eds.), *Cochlear implant and related sciences update* (pp. 198-203). Basel, Switzerland: Karger.

Pisoni, D.B. (1997). Perception of synthetic speech produced by rule: A selective review and interpretation of research over the last 15 years. In J.P.H. van Santen, R.W. Sproat, J.O. Olive, & J. Hirschberg (Eds.), *Progress in speech synthesis* (pp. 541-560). New York: Springer-Verlag.

Pisoni, D.B. (1997). Some thoughts on "normalization" in speech perception. In K. Johnson and J.W. Mullennix (Eds.), *Talker variability in speech processing* (pp. 9-32). San Diego, CA: Academic Press.

Whalen, D.H., & Sheffert, S.M. (1997). Normalization of vowels by breath sounds. In K. Johnson and J.W. Mullennix (Eds.), *Talker variability in speech processing* (pp. 133-144). San Diego, CA: Academic Press.

CONFERENCE PROCEEDINGS PUBLISHED

Akahane-Yamada, R., Tohkura, Y., Bradlow, A.R., and Pisoni, D.B. (1996). Does training in speech perception modify speech production? *Proceedings: Fourth International Conference on Spoken Language Processing* (pp. 606-609). Wilmington, DE: University of Delaware and Alfred I. duPont Institute.

Frisch, S. (1997). Similarity, frequency, and exemplars in phonology. In M.G. Shafto and P. Langley (Eds.), *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society* (p. 924). Mahwah, NJ: Lawrence Earlbaum Associates.

Pisoni, D. B., Saldaña, H. M. & Sheffert, S. M. (1996). Multimodal encoding of speech in memory: A first report. *Proceedings: Fourth International Conference on Spoken Language Processing* (pp. 1664-1667). Wilmington, DE: University of Delaware and Alfred I. duPont Institute.

Wright, R., &. Ladefoged, P. (1997). A phonetic study of Tsou. *Bulletin of the Institute of History and Philology, Academia Sinica, 68,* 987-1028.

MANUSCRIPTS ACCEPTED FOR PUBLICATION (IN PRESS)

Akahane-Yamada, R., Tohkura, Y., Lively, S.E., Bradlow, A.R. and Pisoni, D.B. (In press). Effects of extended training on English /r/ and /l/ identification by native speakers of Japanese. *Perception & Psychophysics*.

Amos, N.E., and Humes, L.E. (In press). SCAN test-retest reliability for first- and third-grade children. *Journal of Speech, Language, and Hearing Research*

Bilger, R.C., Matthies, M.L., Meyer, T.A., & Griffiths, S.K. (In press). Psychometric equivalence of recorded spondaic words as test items. *Journal of Speech, Language, and Hearing Research.*

Bradlow, A.R. (In press). Review of P.A. Keating (Ed.), *Papers in Laboratory Phonology III: Phonological Structure and Phonetic Form. Phonology* 13-2.

Bradlow, A.R., Akahane-Yamada, R., Pisoni, D.B., & Tohkura, Y. (In press). Training Japanese listeners to identify English /r/ and /l/: Long-term retention of learning in speech perception and production. *Perception & Psychophysics*.

Bradlow, A.R., Nygaard, L.C., & Pisoni, D.B. (In press). Effects of talker, rate and amplitude variation on recognition memory for spoken words. *Perception & Psychophysics.*

Chin, S.B., & Kirk, K.I. (In press). Consonant feature production by children with multichannel cochlear implants, hearing aids, and tactile aids. In S. Waltzman and N. Cohen (Eds.), *Proceedings of the Vth International Cochlear Implant Conference*. New York: Thieme Medical Publishers.

Forrest, K., Nygaard, L., Pisoni, D.B., and Siemers, E. (In press). The effects of speaking rate on word identification in Parkinson's Disease and normal aging. *Journal of Medical Speech-Language Pathology.*

Frisch, S. (In press). Temporally organized lexical representations as phonological units. In M. Broe & J. Pierrehumbert (Eds.), *Papers in Laboratory Phonology*, Volume 5. Cambridge: Cambridge University Press.

Frisch, S. (In press). Review of Speech Lab, PC software on CD-ROM. *GLOT International* 3.4.

GIERUT, J. A., & MORRISETTE, M. L. (In press). Lexical properties in implementation of sound change. *Proceedings of the 22nd Annual Boston University Conference on Language Development*. Somerville, MA: Cascadilla Press.

HUMES, L.E., AMOS, N.E., AND WYNNE, M. (In press). Issues in the assessment of central auditory processing disorders. *Proceedings of the Fourth International Symposium on Childhood Deafness*. Nashville, TN: Bill Wilkerson Press.

KIRK, K.I., PISONI, D.B. AND MIYAMOTO, R.T. (In press). Lexical discrimination by children with cochlear implants: Effects of age at implantation and communication mode. In S. Waltzman and N. Cohen (Eds.), *Proceedings of the Vth International Cochlear Implant Conference*. New York: Thieme Medical Publishers.

MEYER, T.A., & PISONI, D.B. (In press). Some computational analyses of the PBK Test: Effects of frequency and lexical density on word recognition scores. *Ear & Hearing*.

NYGAARD, L.C., & PISONI, D.B. (In press). Talker-specific perceptual learning in spoken word recognition: Preliminary findings and theoretical implications. *Perception & Psychophysics*.

SHEFFERT, S.M. (In press). Format-specificity effects on auditory word priming. *Memory and Cognition*.

SHEFFERT, S.M. (In press). Contributions of surface and conceptual information to recognition memory. *Perception and Psychophysics*.

SIMPSON T.H., AMOS N.E., AND RINTELMANN W.F. (In press). Effects of pre-existing hearing loss on proposed ANSI S12.13 outcomes for characterizing hearing conservation program effectiveness: a follow up investigation. *Journal of the American Academy of Audiology*.

WRIGHT, R., FRISCH, S., & AND PISONI, D.B. (In press). Speech perception. In J.G. Webster (Ed.), *The encyclopedia of electrical and electronics engineering*, Volume 24. New York: J. Wiley & Sons.

BEST COPY AVAILABLE